

In pursuit of usefulness and distraction

Nature has been redesigned. It is to be hoped that this development, and others introduced in this issue, will enhance not only the publication's value but also its capacity to divert.

I nspect those little vertical lines in the 'footer' at the bottom left of this page, on either side of the volume number. They are a purely artistic element, empty of symbolism, pregnant with nothing but the vision of a designer. They are at the functionless end of the spectrum of changes made in a significant redesign of this publication. At the other, functional end is a set of typefaces that are subtly different from their predecessors in their combinations of size, shape and spacing, and all the more readable for that. In between are new uses of colour, changes in style of presentation and we hope, as a result, enhanced accessibility and clarity.

Nature owes its designers, Esterson Lackersteen of London, a debt of thanks not only for their inspiration but also for their tolerance in having their ideas pulled apart and reconstituted during the trials of editorial practicality and prejudice that all redesigns have to undergo. We are also indebted to groups of readers convened in the United States, Japan, the United Kingdom and Germany who provided useful criticisms and, above all, considerable encouragement to proceed with this development.

As well as the new design, this issue sees the inauguration of several enhancements in content. Authors and readers alike will be pleased to see that we are phasing in the incorporation of titles of papers in the reference lists of Letters, Articles, and Review and Progress Articles (see page 233). The only disadvantage to this development is that authors will find editors somewhat less flexible in enforcing our standard limits on the numbers of references. This is yet another step in our drive for ever more reader- and author-friendliness of the original science that we publish, following hard on the heels of the introduction of Methods sections, increased flexibility in the presentation of figures and enhanced

readability of introductory text.

Time and again, surveys of readers highlight the popularity of our News and Views section. On page 207 is the first example of a new and occasional component: the News and Views Feature. The aim here, as is well illustrated by Christof Koch, is to provide a wide-angle snapshot of a set of recent advances that amount to an important stage in the development of a discipline. Articles of this type will provide a useful complement to the regular coverage in News and Views, being commissioned on the same principles of timeliness and authority. Above all, the editorial emphasis will be on interest and accessibility to a wide set of readers, as opposed to the more in-depth and specialized treatment appropriate to Review Articles.

The Book Reviews section also receives a boost. Following the recent introduction of the "at a glance" reviews of specialized tomes (see page 218), this issue sees the first discussion of a book "in retrospect" (page 217). *Nature* readers are, of course, cultivated people who read widely and reflect intelligently and unpredictably on what they have encountered. We know that many of them have the ability to transmit their reflections and thereby stimulate their fellow *Nature* readers. So, from time to time, we will publish such thoughts on books or other materials that may have been published at any time, and may have been inspiring, unusually provoking or even pernicious in their impact.

These developments, along with others over the past year, are part of a continuing process of evolution (although our mutations are, we would claim, less random than occurs in the wild). On that basis, this publication should continue to provide its readers with an ever more welcome distraction from, as well as support for, their everyday tasks. □

A rush to judgement?

No amount of purely scientific argument seems likely to dispel public concern about Gulf War syndrome.

P resident Bill Clinton's Advisory Committee on Gulf War Veterans' Illnesses has agreed to stay in place beyond its initial term of 18 months, keeping an eye on the collection of ailments known as Gulf War syndrome. The panel shows admirable devotion to duty, for it must know that nothing it says or does is likely to satisfy the substantial body of opinion that believes the veterans are suffering because of their experiences in the Gulf, and that government agencies and the medical profession have been covering up ever since.

The advisory committee concludes in its report, published last week (see page 187), that current scientific evidence "does not support a causal link" between the symptoms reported today by Gulf War veterans and exposure to any of the commonly discussed environmental risk factors. In fact, the report suggests that the reported ailments may be little more than the stress-induced symptoms familiar to veterans of every major conflict of this century. But the current political atmosphere in the United States, characterized by a strong distrust of the

federal government, is not conducive to this explanation.

Three studies published this week in the *Journal of the American Medical Association* suggest, on the basis of self-reported information from a small group of veterans, that Gulf War syndrome consists of several neurological syndromes caused by exposure to a combination of chemical weapons, vaccines and insecticides. Despite the apparent limitations of the studies, they were subsequently used at a Senate hearing last week to exploit public sympathy for the veterans and lambast the president's panel for its conservative findings.

There is a danger that the issue could develop a public profile disproportionate to the nature of the problem, as has happened with the 'prisoners of war — missing in action' issue which has fixated many Vietnam War veterans for the past twenty-five years. Gulf War syndrome has yet to establish its medical credibility; it could still prove ultimately of greater interest to social historians than to medical researchers. □



US claims of 'no chemical links' to Gulf War illnesses under fire

[WASHINGTON]. The controversy surrounding claims of a 'Gulf War syndrome' took a new turn last week, when a US presidential panel reported that the ill-defined collection of maladies could not be tied to chemical weapons or other environmental exposures in the Persian Gulf.

The following day, the lead author of three new scientific papers suggested exactly the opposite at a press conference. And the papers' conclusions were immediately used by members of a Senate committee to attack the advisory panel's conclusions as being too conservative.

On 7 January, President Bill Clinton accepted the findings of the Presidential Advisory Committee on Gulf War Veterans' Illnesses, which he had set up in May 1995 to investigate the mysterious afflictions reported by tens of thousands of Gulf War veterans.

Its 126-page report sharply criticized the Pentagon for laxity in investigating troops' possible chemical and biological exposures. But it found no scientific evidence of a causal link between veterans' illnesses and exposures to various environmental agents — including chemical weapons — in the Gulf.

The panel also concluded that stress was likely to be "an important contributing factor" to the range of physical and psychological symptoms reported by veterans. It recommended that federal research on Gulf War illnesses should focus more on the physiological effects of stress.

But the following day, the editor of the *Journal of the American Medical Association (JAMA)* announced that this week's issue of the journal would carry three papers suggesting that exposure to combinations of chemicals — including chemical nerve agents, government-issued insect-repellent, and a drug taken to prevent nerve-gas poisoning — appear to be responsible for specific neurological syndromes in veterans.

"The findings of our study provide, to our knowledge, the first epidemiologic evidence of associations between environmental risk factors and systematically defined syndromes in Gulf War veterans," the authors of the papers write.

At a news conference, Robert Haley, director of epidemiology at the University of Texas Southwestern Medical Center, Dallas, and lead author of the three *JAMA* papers, said: "This will give veterans a lot of heart." Referring to joint and muscle pain, depression, chronic fatigue, skin rashes and chronic diarrhoea reported by Gulf veterans, Haley said that his work "probably explains all these unusual symptoms".

The papers' findings were immediately picked up by politicians from both political parties, who used them at a Senate hearing on 9 January to attack the president's committee for being too conservative both in its report and in its reaction to the new studies. There is "a lot of evidence" about a causal connection, said Arlen Specter (Republican, Pennsylvania), the new chairman of the Senate Veterans' Affairs Committee. "We need to pinpoint it."

But Philip Landrigan, director of environmental and occupational medicine at Mount Sinai School of Medicine in New York, and a member of the presidential panel, argued at the hearing that the selection methods and relatively small sample size (23) in one of the *JAMA* studies made it impossible to generalize its findings to the 697,000 US Gulf War veterans. The papers "don't prove the point," said Landrigan. "They raise the question. They suggest the urgent need for further research."

Landrigan's arguments also appear in an accompanying *JAMA* editorial in which he writes that the Haley studies "have limitations that substantially weaken the authors' strong conclusions". This did not prevent Jay Rockefeller (Democrat, West Virginia), the senior committee Democrat, from lambasting the presidential panel for an "overriding focus" on stress as a source of post-Gulf War illnesses.

He called this a "wonderful excuse" for the government to ignore chemical agents that could be the source of illness. He suggested that independent studies — as opposed to those sponsored by the US government — were finding chemical causes for the illnesses afflicting veterans.

The chair of the presidential committee, Joyce Lashof, a former dean of the school of public health at the University of California at Berkeley, immediately challenged Rockefeller. "Not all independent research goes in that direction," she said. She added that her committee's conclusions were drawn from a review of both government-sponsored and independent research.

The first *JAMA* paper uses factor analysis to derive various neurological syndromes, which were found among 25 per cent of the veterans studied. The second documents diffuse neurological abnormalities among 23 veterans with one of three primary syndromes. The third links the syndromes to different chemical exposures.

'Impaired cognition syndrome' (including difficulty remembering, depression and insomnia) occurred more frequently in

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Fighting for breath? Stress, rather than chemical agents, may have caused 'Gulf War syndrome'.

subjects who reported wearing flea and tick collars containing the pesticide chlorpyrifos.

'Confusion-ataxia syndrome' (thinking, reasoning, and balance problems) occurred eight times more often among veterans who believed they had been exposed to chemical weapons. This effect was synergistically amplified in veterans who had serious adverse reactions to the anti-nerve gas medication pyridostigmine bromide (PB).

In addition, 'arthro-myo-neuropathy syndrome' (joint and muscle pain) occurred more frequently with increasing use of government-issued insect-repellent containing the chemical DEET.

The researchers say that the neurological syndromes are "best explained" as variants of a syndrome of nerve damage called organophosphate-induced delayed polyneuropathy "resulting from exposure to combinations of organophosphates and other cholinesterase-inhibiting chemicals, including nerve agents, pesticides, insect repellents and pyridostigmine bromide".

The presidential committee recommended research on the long-term effects of low-level exposure to chemical nerve agents, and the synergistic effects of exposure to PB with other Gulf War risk factors. It called for an independent investigation of troops' possible exposure to chemical and biological warfare agents, and criticized Pentagon investigations as "superficial".

Government departments have been given 60 days to draw up a plan for implementing the committee's recommendations. The committee's life has been extended for nine months, allowing it to oversee the implementation of its recommendations, and to oversee a Pentagon investigation of possible biological and chemical exposures of Gulf War troops.

Meredith Wadman

Rio review to rejuvenate green initiatives

[WASHINGTON] Environmentalists are planning to use this year's fifth anniversary of the Earth Summit that took place in Rio de Janeiro in June 1992 to attempt to regain some of the momentum which has been lost since the meeting closed.

Up to 500 representatives of governments, environmental groups and the private sector will meet in Rio in the third week of March for an event called RIO+5 to review the successes and failures of the years since the 1992 summit.

Two months later, the general assembly of the United Nations is due to meet in special session to see what progress member states have made in implementing Agenda 21, the document they signed at Rio.

Maurice Strong, the Canadian industrialist and veteran environmentalist who served as secretary-general of the Earth Summit, said in Washington DC last week

that RIO+5 is "designed to regenerate some of the momentum, which, to some degree, has faltered" since 1992.

According to Strong, although many good things have happened as a result of the Rio meeting, "the basic trends in our economic life which have caused the environmental problems have not changed, and the environment continues to deteriorate".

The track records of governments, particularly in developed countries, have come under attack in two reports published last week. One is produced by the Earth Council, an international pressure group which Strong chairs, and the other by the Worldwatch Institute, based in Washington.

The environmentalists argue that Rio has had some positive impact, pushing some governments and corporations to listen to green arguments at various forums established since the summit, and to incor-

porate some of the issues raised in their decision-making.

Strong cites the setting up of more than a hundred national councils for sustainable development as one of Rio's principal achievements, although conceding that their performance has been "mixed". He also says that some 1,500 cities across the world have set up committees to monitor local compliance with Agenda 21.

Environmentalists claim that the developed countries have failed to live up to promises made in Rio. The United States, for example, has failed to ratify the Biodiversity Treaty, cut back on aid that developing countries said they would need to comply with Agenda 21, and is unlikely to meet the Rio goal of stabilizing emissions of greenhouse gases at 1990 levels by the end of the decade.

Strong was restrained in his criticism of the governments of these countries, arguing that the public had to support painful changes, such as higher fuel prices, before such changes could be implemented. "Even Canadians won't accept increases in energy prices," he concedes. "I am frustrated, but the moment you give up, your pessimism will become self-fulfilling."

Jonathan Lash, president of the World Resources Institute (another Washington-based environmental group) and co-chair of the President's Council for Sustainable Development in the United States, says that despite its obvious failings, Rio has achieved a notable list of successes.

The summit established the environment as "an important part of the dialogue of international leaders" for the first time, he says. It had encouraged the private sector to take environmental issues seriously, and environmentalists to consider economics and human needs in their campaigns. And the summit confirmed the establishment of a new source of development funds, the Global Environmental Facility (see *Nature* 385, 106; 1997).

But as the Earth Council points out in its report, the major problems addressed at Rio have continued to fester. The council points out that carbon dioxide emissions have grown in developed and developing countries, while deforestation and biodiversity losses continue apace.

The Worldwatch Institute's annual State of the World review says that, in its scale and scope, the summit "set a standard for itself that was almost certain to lead to disappointment". It calls for the establishment of an "E8" of the countries that "shape global environmental trends" — China, the United States, Brazil, Germany, Japan, India, Indonesia and Russia — to "catalyse action" on the environment, along the lines of the G7 group of capitalist economies.

Colin Macilwain

'Industrial apathy' hampers climate schemes

[NEW DELHI] The prospects of slowing down climate change through a scheme in which industrialized countries sponsor the reduction of greenhouse gas emissions in the developing world appear to be receding. That was the conclusion of a meeting organized by developing countries in New Delhi last week.

Representatives of 150 countries agreed that idea of 'joint implementation' is sound, but reported slow progress. Only 30 projects have been approved since the pilot phase was launched two years ago.

Joint implementation was promoted at the 1992 Earth Summit (see above). The idea seemed attractive, as it offers advantages to both developing and industrialized countries.

Greenhouse gas emissions from industrializing countries are likely to rise sharply in the next decade. Under joint implementation, developed countries transfer low-cost energy-efficient technologies to the developing world.

Developing countries saw an opportunity to obtain private capital to finance new

technologies. In return, credit for energy saved is passed on to developed countries as part of their greenhouse gas reduction targets.

The terms under which such a strategy would work have yet to be established. But parties to the UN climate convention had hoped that these issues would be resolved during the pilot phase of carbon-reducing activities implemented jointly.

"It is a matter of concern, given that most of the projects have not reached the operational phase," according to a statement agreed by participants at the conference.

A consensus emerged at the conference that the 2000 deadline for evaluating the pilot phase is too close, and that information about projects is unlikely "to serve as a credible basis for a decision on whether to move forward". Some felt that the pilot phase may have to be extended by five or ten years.

The lack of pilot projects is also being attributed to industrial apathy. The conference heard that neither new technologies nor private investment for project finance have materialized.

The prospect of significant participation from business remains remote, particularly as investment in carbon-reducing projects is unlikely to yield significant returns. "[A pilot] project is like an R&D project, and it has potential to get profit in future," says Hidehisa Tanaka of Mitsubishi Research Institute of Japan. In his view, companies will not participate unless projects "are formulated from a commercial base and some incentives are provided".

The pilot phase is also the victim of a lack of trust between developed and developing countries. A delegate from Zimbabwe pointed to "a tendency of the investor countries to target host locations where they can maximize commercial profits".

But Ashok Khosla, who organized the conference on behalf of Development Alternatives, an Indian nongovernmental organization, believes that "all hope is not lost", as many potential pilot projects may be converted into actual projects in the three years before the pilot phase comes up for review.

K. S. Jayaraman

Investment analysts paint rosy prospect for biotech

[SAN FRANCISCO] After a sluggish stock performance in the second half of 1996, the US biotechnology industry is poised for unprecedented growth in 1997, according to stock analysts who point to the large number of new drugs being developed, strong finances and an improved regulatory environment.

Persuaded by this optimistic assessment, investors at the annual Hambrecht & Quist Health Care Conference in San Francisco last week were hastily buying shares in the companies that announced details of 200 drugs in late-stage clinical trials.

The data on those treatments are likely to be released during the next 18 months. Regeneron, Alpha-Beta, AutoImmune, Creative Biomolecules, BioChem Pharma and ImmunLogic are likely to be among the first to report on important trials.

Most companies are in a strong position to build on the results, according to Richard van den Broek, a biotechnology analyst at Hambrecht & Quist (H&Q). The sector as a whole has recently raised US\$5 billion, and the average company has three years' supply of cash in its coffers, he said.

Analysts argue that drug developers are enjoying an improved relationship with the US Food and Drug Administration (FDA), which has cut drug approval times dramatically over the past two years. They also say that with the FDA commissioner, David Kessler, stepping down, biotechnology companies expect the changes to lose some momentum, but not to go into reverse.

The upbeat assessment is also based on the fact that some industry experts are

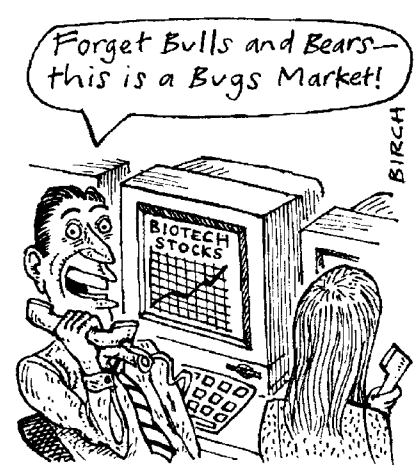
expecting proposals to loosen standards of efficacy to get drugs to the market sooner. At the same time, some of the potential threats that had been feared by drug developers over the past year have not materialized.

A year ago, for example, company executives were concerned about the potential impact of managed care on the acceptance of innovative treatments. Managed-care companies are now moderating their emphasis on cost and demonstrating a greater interest in quality, said Robert Faulker, a medical devices analyst at H&Q. This has helped development of drugs and devices, particularly those that had cost-benefit analyses built into their clinical trials.

Although companies are building on a strong year, investors did not do so well in 1996. After a profitable first half, biotechnology investments in the second half of the year lagged behind the broader market. Despite this, Dennis Purcell, managing director for life sciences investment banking at H&Q, predicted healthy growth in company valuations and stock market returns for 1997.

Purcell pointed out that, over the first two days of the conference, 85 per cent of the companies that made presentations enjoyed stock price increases, even though the market was flat. One company, Geron, almost doubled in price. "1997 may be the year of the investor," said Purcell, who depicted biotechnology as a promising investment target.

Rachel Leheny, a biotechnology analyst with H&Q, argued that the industry appears to be entering an era of efficiency, in terms of both science and business management.



Important technology 'platforms', such as genomics, screening and combinatorial chemistry, are now in place, providing the means for rapid identification and development of well-targeted drugs. Not only are development times likely to improve, but the resulting drugs are likely to be more effective, she said. The result will be fewer clinical failures and a dramatic improvement over the pharmaceutical industry's current 10 per cent success rate in drug development.

According to Steven Burrill of the merchant bank Burrill & Company, European companies are leading the way to programme swaps and other strategic links between companies. He said that Europe has become a hotbed of young companies as technology transfer out of the academic environment has accelerated. Purcell suggests that US companies may go to Europe to gain access to capital and to take advantage of patient and hospital systems there.

A flow of new discoveries has increased the attraction of the biotechnology industry to scientists, many of whom are moving to smaller companies from the major pharmaceutical corporations.

Sally Lehrman

Glaxo Wellcome turns to worms for guidance on genes

[SAN FRANCISCO] Just three months after acquiring a subsidiary that specializes in nematode genetics, Sequana Therapeutics of La Jolla, California, has signed a research agreement with Glaxo Wellcome to use its technology for functional analysis of human disease genes.

According to Tim Harris, senior vice-president and chief technical officer at Sequana, the move reflects the fact that large pharmaceutical companies are beginning to recognize the value of less complex organisms for understanding the function of human genes, identifying targets for therapeutic approaches and getting them into screening.

Under the agreement, NemaPharm of Cambridge, Massachusetts, will use the nematode worm *Caenorhabditis elegans* as a model animal system to evaluate genes, characterize the pathways of their activity and identify targets for therapeutics. The exact genes and their disease areas have not been disclosed.

Neither have the financial terms of the agreement been revealed, although Glaxo Wellcome says it will provide full research support and make payments when NemaPharm identifies targets that become part of Glaxo Wellcome's drug-

discovery programmes.

NemaPharm is also to receive royalties on products that result.

The nematode is expected to be fully sequenced by 1998. Of the 70 human disease genes cloned so far, more than half have known homologues in the worm. In a presentation at the Hambrecht & Quist Health Care Conference in San Francisco last week (see above), Kevin Kinsella, chief executive of Sequana, emphasized the animal's well-understood biology and the ease with which its genes can be manipulated.

Among other things, NemaPharm has used the model

to find a chemical that can block apoptosis, and to knock out a ras gene homologue and return its function to normal, Kinsella said. The company will do about one new gene knockout a week.

Sequana acquired NemaPharm in October. The company was set up in 1990 in Cambridge, Massachusetts, with founders that included Robert Horvitz, professor of biology at the Massachusetts Institute of Technology.

Along with computational biology, the work of the subsidiary has since become crucial to Sequana's functional genomics programme, said Harris.

SL

Embryo researcher is sanctioned for using NIH resources

[WASHINGTON] The US National Institutes of Health (NIH) confirmed last week that it has cut all research links with a leading reproductive biologist for breaching a ban on the use of federal funds for research on human embryos.

According to Anne Thomas, spokeswoman for NIH director Harold Varmus, the biomedical agency ended a contract with Mark Hughes, director of Georgetown University's Institute for Molecular and Human Genetics in Washington DC, last October.

"What concerned us was that [Hughes] used NIH equipment and government-funded research fellows to do pre-implantation diagnosis," says Thomas. This was despite the fact that he had been told that NIH resources could not be used in this manner.

Hughes had been working at the National Center for Human Genome Research (NCHGR) under a contract arrangement by which part of his salary was paid by Georgetown with funding received from NIH.

The work that caused concern was not conducted at NCHGR, where Hughes, an expert in single-cell biology, was doing experiments that included creating cDNA libraries from a single cell. Thomas says one or more postdoctoral fellows complained to NCHGR officials about separate research on which Hughes had asked them to collaborate at the Montgomery Fertility Institute, a private practice at Suburban Hospital, near NIH in Bethesda, Maryland.

Hughes had been conducting pre-implantation diagnosis on test-tube embryos at the private institute, removing a single cell at the eight-cell stage to screen for genetic abnormalities and implanting the remaining seven cells if none was found.

The work is not illegal, and is done privately in the United States, the United Kingdom and elsewhere. But Congress banned federal funding for human embryo research in 1996, and renewed the ban for the fiscal year 1997, which ends on 30 September.

Neither Hughes, Georgetown University nor NCHGR officials were available for comment this week. Earlier, Hughes told *The Washington Post* that "there has never been any intention of doing anything wrong". The newspaper said that he regretted having used government resources at the private clinic, but made no apologies for the work itself.

In the past, Hughes has opposed the federal ban (see *Nature* 376, 288; 1995). He was a member of an expert NIH panel which in 1994 concluded that human embryo research was acceptable — within strict constraints — before the primitive streak appears, at day 14. **Meredith Wadman**

Russian space module 'in need of more protection'

[WASHINGTON] Russian-built sections of the international space station will have to be fitted with extra shielding to protect against damage from orbital debris, says a report published last week by the US National Research Council (NRC).

The NRC, the operational arm of the National Academy of Sciences, says that the station may receive warnings about possible collisions with large objects more often than previously thought — ten times a year instead of six times. But it adds that the National Aeronautics and Space Administration (NASA) and its European, Japanese and Canadian partners have a sensible plan to counter the threat from debris.

Shields on the station's outer surfaces will pulverize small objects that strike the station at high speed. But Russian components under construction — particularly a 'service module' that will house the initial crew of three — have not been shielded to the same standard as other sections of the station.

The risk is not that the pressurized module would be penetrated by debris — the walls are thick enough to prevent that — but that plumbing and other exposed hardware on the outside would be damaged, according to Nicholas Johnson of the NASA Johnson Space Center. He was originally a member of the NRC panel, but resigned earlier this year when he took a job as NASA's expert on space debris.

Even before the NRC report appeared, US and Russian engineers had a solution in mind. Spacewalking astronauts could fit the Russian components with additional shielding after the hardware is delivered to orbit.

Installing the shielding before launch is

ruled out by the way the module fits on top of its booster rocket, says Johnson. But the proposed strategy has the advantage of deferring any expense associated with shielding to later in the programme.

Russian engineers have tended to play down the threat from debris, pointing out the lack of serious problems in their decades of experience with small space stations. Johnson agrees that NASA errs on the side of caution: "We have placed high standards on the space station in terms of reliability."

Objects larger than 10 cm in diameter could cause serious damage if they hit even shielded parts of the station, so the strategy for them is different: get out of the way. US military surveillance cameras that track orbiting junk would issue a warning whenever an object was about to come too close to the station, and the entire structure would be moved.

Such warnings have been issued in the past for the US space shuttle, and recently the Americans began providing the same service to the Russian Mir space station. The shuttle has been moved out of the way to avoid a collision, but Russian space officials have so far chosen not to move Mir on the few occasions it has been warned, trusting in the very long odds against a collision.

Even if the number of warnings does increase from six to ten a year — as the NRC report suggests that it might, on the basis of recent predictive models — international space station managers would also be free to ignore a warning. They might do so if some delicate, ultra-quiet microgravity experiment would be jeopardized by firing the station's manoeuvring engines. **Tony Reichhardt**

Cassini gets golden gown made to measure

NASA/JPL

[LONDON]. The art of bespoke tailoring was called upon last month to cut, stitch and fit a shiny, gold-coloured protective suit for the Cassini spacecraft, due to begin a seven-year voyage to Saturn in October.

Technicians and engineers working for the NASA used traditional industrial sewing machines, brown paper patterns and large cutting tables to make the thermal blankets to protect Cassini from the extreme temperatures of deep space.

Mark Duran, supervisor of NASA's 'shield shop', which provides survival gear for spacecraft and instruments, says that Cassini's super-strong and extremely lightweight blankets "are built unlike any others". They are designed to withstand temperatures from -220 to $+250$ °C.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

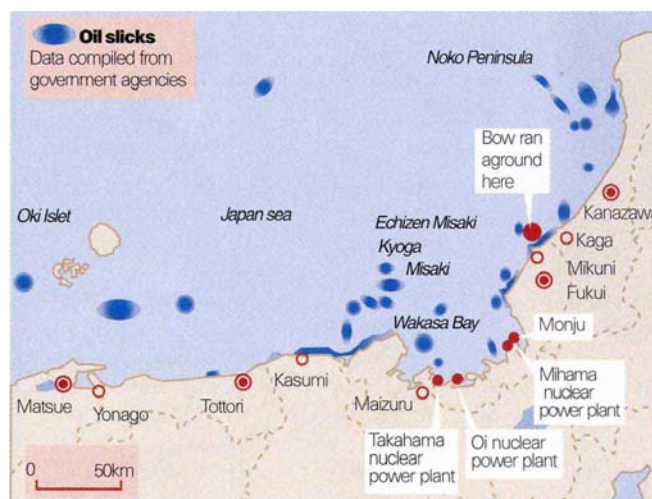
Oil spill threatens Japanese nuclear plants

[TOKYO] Japan's worst-ever oil spill is threatening to shut down nuclear power plants along the coastline of the Japan Sea as large oil slicks approach the seawater intakes of the stations. The spill was caused when the Russian oil tanker *Nakhodka* broke in two 110 km off the Oki Islands during a storm on 2 January. Millions of litres of thick fuel oil have leaked from the tanker.

Russia has despatched an oil-cleaning ship from Sakhalin to help in the clean-up operation. But, as with previous disasters such as the Kobe earthquake, the central and local governments in Japan are attracting severe public criticism for their slowness to respond to the disaster. Most of the clean-up effort so far has been by local people and volunteers.

One half of the tanker sank in waters 1,700 metres deep and cannot be salvaged. The bow section washed up on the coast of Fukui Prefecture near the city of Kanazawa, and is being buffeted by high seas that are hampering salvage operations and attempts to contain the leaking fuel. Both sections still contain millions of litres of oil.

Small slicks have washed up at dozens of locations along the coast, and thousands of local people and volunteers from around Japan have been working frantically to clean



Danger zone: the strip of coastline 400 kilometres to the west of Tokyo that is currently threatened by oil slicks. The oil resulted from the breaking-up of a Russian tanker that was caught in a storm at the beginning of January. Large slicks are nearing Wakasa Bay, where several nuclear power plants are located.

up the spills with buckets and shovels. A pit has been prepared to temporarily hold 2.7 million litres of oil from the tanker which was carrying 19 million litres when it sank.

But clean-up operations at sea have been hampered by rough conditions, making it difficult for workers to fight slicks with oil-dispersing chemicals, oil mats and oil fences. A 300-metre-long fence set up around the bow section of the tanker by the Maritime Safety Agency was quickly torn apart by rough seas.

Large slicks more than 10 km across are approaching Wakasa Bay where there are several nuclear power plants, including the Monju fast-breeder reactor (see map). The water intakes of the plants are protected by oil fences. But it is doubtful whether they will be sufficient to keep the slicks at bay. If any oil were to enter the intakes, which suck in vast amounts of sea water, it could seriously damage the water-cooling systems for the turbines and lead to the close-down of the plants.

David Swinbanks

Australia maintains commitment to joint research centres

[SYDNEY] Australia's conservative Coalition government, which came to power last March, has allocated its first funding to the country's Cooperative Research Centre (CRC) programme. The scheme was introduced by the former Labor government in 1990 in a bid to stimulate closer collaboration between universities, government research agencies and industry. Sixty-one centres have been funded since the start of the programme.

The latest funding round included the first scheduled reviews of 13 centres five years after their establishment. Each centre was seeking renewal of funding beyond its initial seven years. Out of 34 applications, 16 — made up of ten renewals, five new centres and one extension — were approved by Peter McGauran, science and technology minister, who last week allocated A\$218 million (US\$173 million) to the 16 centres over seven years.

This means that the government has preserved — in spite of budget restrictions — the overall level of annual funding established by its predecessor of around A\$145 million a year for the CRC programme.

The new centres will be established in the fields of Aboriginal and tropical health,

molecular plant breeding, sustainable rice production, sustainable tourism and the discovery of genes for common human ailments.

Two of the existing CRCs had decided not to reapply for funding, and three — tropical pest management, robust and adaptive (computer) systems, and plant science — failed to win renewed contracts.

The Plant Science CRC, established in Canberra jointly by the Australian National University and the division of plant industry of the Commonwealth Scientific and Industrial Research Organisation (CSIRO), had 27 industry associates, and support from industry had grown rapidly to more than one-third of its budget.

Jim Peacock, co-director of the centre, says it will reapply for government support, but this cannot be done until 1998 when current funding ceases. "While we are pressing ahead with the research projects which our industry partners want, the decision [not to renew the contract] is very disappointing and affects our continuity as we have to shed research staff."

Mark Sceats, director of the Photonics CRC and chairman of the body that represents all centre directors, points out that new research opportunities and jobs are

emerging from what he describes as a "beneficial change in research culture".

Sceats says that the closure of some centres and opening of others is "a natural, organic part of the CRC concept". But he expresses concern that the government has erected "very high barriers" to involvement by industries in the programme by asking for a seven-year commitment.

He believes that, with five-year reviews, CRCs are being expected to deliver results two to four times faster than the normal timescale of 10 to 20 years for "moving innovation from our universities and CSIRO through to products and services which generate economic benefits".

At the end of 1995, Labor's Innovation Statement made a commitment to three additional centres (see *Nature* 378, 653; 1995), but these appear to have been quietly dropped.

Total funding of the centres represents about 4 per cent of the Australian government's expenditure on research and development. On average, the value of contributions in cash or kind from academic, government and industry partners in an individual centre amounts to two or three times the government's cash support.

Peter Pockley

Europe seeks to rally regional research

Alison Abbott

As political leaders fight about the economic implications of introducing a common currency, the European Commission in Brussels is grappling with similar problems over the support of research.

[BRUSSELS] The promotion of research activities in Europe's poorest regions using so-called 'structural funds' is growing — and being closely scrutinized. As with the currency, at the root of the problem lies the tension between policies that favour those able to use financial resources most productively, and others designed to raise the capabilities of Europe's poorer, less productive, regions.

Cohesion between regions

The European Union (EU) aims to achieve 'cohesion' by ironing out economic differences between regions. As part of this strategy, the commission will spend nearly ECU11 billion (US\$8.8 billion) to help build the research bases of poorer regions by the end of the century.

But the effectiveness of this funding is uncertain, so officials responsible for allocating money directly to research through the separate Framework programme are reluctant to give poorer countries special treatment. The strain this creates will be addressed in a 'communication' from the commission in April.

According to a report on cohesion policies published by the commission last month, there is some evidence that heavy investment may be beginning to pay off — the gap in the research capabilities of rich and poor countries is slowly narrowing.

But the report criticizes an over-emphasis on research excellence in the commission's Framework research programmes, which it says has led to the best academic

institutes in the north clubbing together with the best institutes in poor countries. In the long run, it says, this could distort the research agenda of poorer countries which need to develop applied research capabilities more relevant to their needs.

The report exposes tensions between the EU's mandate of promoting the industrial competitiveness of the union as a whole and that of selectively promoting poorer regions. The first means funding the best research through Framework programmes, the second supporting that in less developed regions through subsidies.

The April communication will attempt to smooth out these tensions, particularly after 1999, when the current round of 'structural funds' — subsidies to poorer regions — and the current Framework research programme expire. In response to pressure from the United Kingdom and Germany, the commission is likely to recommend a move away from the strategy of basing subsidies primarily on geographical location, towards one that addresses issues common to all member states, such as urban decline, rural regeneration and youth employment.

This would be in line with the commission's plans for its fifth Framework research programme, which stress the importance of social needs (see *Nature* 382, 194; 1996). Standards of living vary fourfold between EU countries, and efforts to narrow the gap consume one-third of the commission's budget.

The technology gap is even bigger. Portugal, for example, spent only 0.63 per cent of its gross domestic product on research in 1992, compared to 2.48 per cent in Germany and nearly 3 per cent in Sweden. Portugal has less than a quarter of the number of researchers in Germany, measured as a percentage of the labour force. It also submits 11 times fewer European patent applications, as a function of total research investment, than Germany.

Sharp differences can also occur between regions of the same country. Within Italy, the north has seven times as many researchers as the south, in proportion to population.

Richer countries, particularly the United Kingdom and Germany, are unhappy about the size of the EU's structural funds, as well as the way they are used. In the interests of

'subsidiarity' — the union's principle that decisions should not be made centrally when they could be made locally — structural funds for poor countries or regions are allocated by the recipient countries themselves.

The biggest recipients include Ireland, Spain, Portugal and Greece — the four 'cohesion countries' — and the regions of southern Italy and eastern Germany. The commission would like them to spend more on long-term investment, including research, to reduce the technology gap. But their politicians tend to prefer shorter-term, vote-catching measures, such as motorway building.

Despite subsidiarity, the commission has managed to exert some influence. With the small percentage of structural funds over which it has direct control, it launched in 1990 its ECU400-million Stride programme (Science and Technology for Regional Innovation and Development in Europe). This has supported more than 1,000 projects, and in 1994 research and development was made a priority in its guidelines for structural fund spending.

Such measures have already borne fruit. Before 1987, only ECU180 million of structural funds had been spent on research. This rose to ECU3.2 billion between 1988 and 1993, and an additional ECU7 billion is to be spent on research and development in the latest spending round (1994–99).

The 'cohesion four' have increased the proportion of their structural fund allocation spent on research from 4 to 6 per cent. Most has been spent on infrastructure and equipment. But countries are increasingly willing to spend the money on training researchers (see figure, below left).

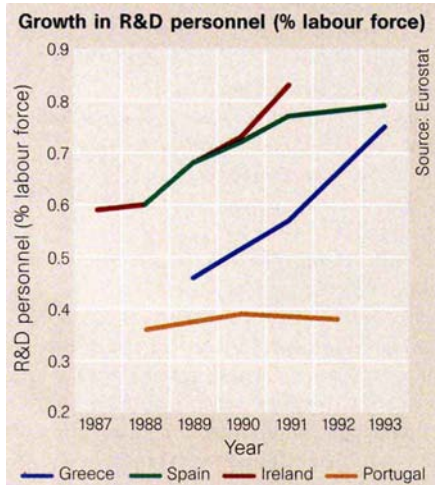
Rule change needed?

Eneko Landaburu, director general of the commission's directorate for regional policy and cohesion, says the improvements do not go far enough. He would like to see the rules on structural funds changed to enable more to be spent on research and development.

Politicians and researchers have repeatedly called for such a move. The European Science and Technology Assembly (ESTA), a group of 100 academic and industrial scientists that advises Edith Cresson, the research commissioner, has stressed the need to use structural funds to strengthen a country's research capability.

That would allow poor regions to compete on equal terms with richer countries for money from the Framework programme, rather than seeking preferential treatment from the research commissioner. "Researchers don't like the idea of structural funds being given to countries which are then left free to spend them how they like," says Jan Borgman, ESTA's chairman.

Even the commission cannot tell whether structural funds officially used to build up



research capabilities have been spent sensibly. No detailed information has been collated on how the money has been spent, and officials say that the long-term nature of such investment makes it too early to judge whether it has been a success.

But there are signs that the technology gap has been slowly narrowing. National expenditure on research and numbers of research workers per capita in the four cohesion countries have grown since 1989. So too has scientists' publication output.

Tom Higgins is director of the Dublin-based consultancy Circa, which has analysed the use of the 1988-93 structural funds for research in Portugal, Ireland and Greece. He says this improvement "is likely due to structural funds, simply because of the large amounts of money involved".

But Higgins says that the commission "should not assume that such parameters indicate that the primary goal of cohesion policy — increased industrial competitiveness — has been achieved". He says there has been neglect of the 'demand' side of research effort — local industrial needs — while too much has been spent on 'supply', the building of infrastructure. The commission concedes this point.

Most structural funds relating to research have been spent on infrastructure. For example, 50 new research institutes have been built with structural funds in Portugal and 100 have been refurbished.

Building to a critical mass

Anecdotal evidence suggests that new research institutes in scientifically remote regions are most likely to be successful if they are backed by experienced research institutes in richer areas, or can quickly build up critical mass of highly trained researchers (see panel, right).

But spending large sums on research infrastructure is not as easy as it sounds. Structural funds are withdrawn if not spent in the year allocated, so it is easier to attach them to projects already planned.

The creation of a marine research institute at the University of Galway in western Ireland in 1993 is one example of a planned project that was awaiting only financing. New ideas for projects are harder to get started fast enough. The Stride programme, originally intended to run for only four years, has been extended by three years because of the difficulties some countries have in rapidly using large amounts of funding.

Italy has spent only 70 per cent of its Stride money and is unlikely to be able to spend the rest before the programme ends in December. It has had the biggest problems, partly because its political instability has upset the legal structure needed to administer the money.

A desire to make the most of an opportunity that may soon end partly explains

Italy's interest in proposing itself as a possible location for ITER, the next-generation experimental fusion reactor. Structural funds would pay for the infrastructure (see *Nature* 384, 507; 1996).

A further indicator of the success of cohesion policies is the increased participation of the four cohesion countries in the Framework research programme.

Between 1987 and 1990, half the Framework funds went to scientists working in 12 so-called 'islands of innovation' which stretch between London and Milan and include Amsterdam/Rotterdam, Ile de France

and Munich. This is despite the fact that what is nicknamed Archipelago Europe accounts for only 28 per cent of the EU population. In the same period, only 10.5 per cent of funds went to the four cohesion countries. Between 1991 and 1994 the share of the 'archipelago' had fallen to 45 per cent, while that of the cohesion four had risen to more than 12 per cent.

The commission's directorate for cohesion believes Framework is still elitist. But any pressure to make the fifth Framework programme more favourable to cohesion regions is unlikely to win through. □

Molecular biology reaps the benefits

[MUNICH] Those seeking evidence of the success of structural funding need look no further than the Greek Institute of Molecular Biology and Biotechnology (IMBB). Established in 1984 in Heraklion on the island of Crete, the institute has become a flagship of European research.

The institute was built and equipped almost entirely with structural funds, which have also supported some of its projects in the 1990s. It has a staff of 150, a graduate school programme and a strong publication record.

When the Greek government proposed the use of European Union (EU) money to establish a series of research institutes outside Athens in the early 1980s, Fotis Kafatos, then professor of molecular biology at Harvard University in the United States, returned to his native Greece to become IMBB's first director.

By 1990 the institute was participating in 12 projects of the European Commission's Framework research programme, where scientists from different countries must join forces to compete for funds.

By the end of 1992, scientists at the institute were running 30 programmes and they now participate in 45.

The institute was the first to demonstrate transfer of genes into an insect other than the fruitfly *Drosophila*. The recipient was the Mediterranean fruitfly *Ceratitis*

capitata, a species of agricultural importance. Structural funds financed the whole project.

One of the secrets of the institute's success, according to Kafatos, who left in 1993 to become director of the Heidelberg-based European Molecular Biology Laboratory, was the opportunity given by

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Thireos: regrets lack of interest in research from local industry.

the generous structural funds "to bring in large numbers of highly qualified, foreign-trained scientists and give them the appropriate conditions in which to work".

It was also important, he says, that the institute was not built in isolation but as part of a complex of research institutes belonging to Greece's Foundation for Research and Technology.

The IMBB's current director, George Thireos, is proud that after "wisely spending" structural funds, the institute is now able to survive on competitive grants. But he regrets that plans to spin off the institute's enterprises — which last year brought in ECU112,000 — have been hindered by local

industry's lack of interest in research.

Italy's Mario Negri Sud research institute at Santa Maria Imbaro was one of 14 institutes set up in the 1980s in Italy's poor south, where the government wanted to stimulate the economy. Even now, only 7 per cent of Italy's national expenditure on research and development is invested in the area.

The institute is the only one of the 14 that continues to do research, mainly in cardiovascular biology. Its success is largely due to the expertise of the Mario Negri Institute for Pharmacological Research in Milan, which helped to set it up, and continues to work with it in close collaboration. The other 13 institutes offer only scientific services.

Despite the bold programme in the 1980s to stimulate development in the south, the institute has been left by the Italian government to sink or swim. Silvio Garattini, director of the Milan institute, regrets the lack of government interest in helping to sustain the EU structural fund investment.

Mario Negri Sud's director, Giovanni de Gaetano, a southerner who trained at the Milan Mario Negri Institute, says structural funds have been well spent. "It has also been comparatively cheap", he says. "The £40 billion (US\$26 million) invested [from structural funds] is equivalent to one or two kilometres of motorway."

AA

French physicians charged over growth hormone 'poisoning'

[PARIS] Five French physicians and health officials who were charged with involuntary homicide in 1993 in the contaminated human growth hormone affair now risk being tried on the more serious charge of poisoning. This follows revelations that the Paris hospital authority continued to distribute potentially contaminated hormone after its use had been forbidden.

The United States banned the use of pituitary growth hormone extracted from cadavers in April 1985 after a confirmed case of Creutzfeldt-Jakob disease (CJD). The United Kingdom imposed a similar ban the following month. France Hypophyse (France Pituitary), the body responsible for producing and distributing the hormone, continued to allow use of the hormone provided it had been treated with urea, a procedure considered to inactivate the prion.

But the magazine *L'Express* revealed last week that the judge investigating the French cases has found that the central pharmacy of the Paris Hospital Authority continued to distribute 20,000 ampoules of potentially contaminated hormone until the early months of 1996. The doses are estimated to have had a value of US\$1 million.

Data 'adequate' for UK nuclear waste lab

[LONDON] Enough is known about hydrogeological conditions at Sellafield in Cumbria to allow the building of an underground laboratory to investigate the storage of low- and intermediate-level nuclear waste, according to scientists asked by the UK nuclear industry's waste-disposal agency, Nirex, to review its research.

Nirex's plans for the laboratory have met with concern about a perceived lack of hydrogeological data at the proposed site (see *Nature* 383, 751; 1996). But the scientists reviewing Nirex's findings have concluded that there is sufficient data both to model groundwater flow in undisturbed conditions and to act as a 'baseline' against which disturbance caused by constructing the laboratory could be measured.

Whale makes waves with record swim

[LONDON] A humpback whale has slashed more than one month off the record for the fastest whale migration from Alaska to Hawaii. The whale, known simply as 339, completed the 2,775-mile journey across the Pacific Ocean in 39 days. The average migration time is about 102 days, with fast migrations at about 80 days, according to

Janet Straley of the University of Alaska South East, who reports the finding in the journal *Marine Mammal Science*.

US gene therapy group to promote research

[WASHINGTON] A group of medical geneticists in the United States has formed an organization to exchange ideas and information about gene therapy, and to promote research and public education. The American Society of Gene Therapy is headed by George Stamatoyannopoulos, professor of medicine and genetics at the University of Washington School of Medicine in Seattle. Its first meeting will take place in Seattle.

Test telescope opens at the South Pole

[SYDNEY] The first astronomical facility from Australia to be set up on the high plateau of Antarctica, which makes up most of the centre of the continent, was opened last week at the South Pole, 2,385 metres above sea level.

The Automated Astronomical Site Testing Observatory is a A\$1-million (US\$790,000) joint venture of the University of New South Wales and the Australian National University, in collaboration with the US Center for Astrophysical Research in



So far, systems in the ÄKTAdesign family include:

- ÄKTaexplorer, for method development and scale up of every biomolecule
- the new ÄKTApurifier, for purification of peptides, oligonucleotides and other biomolecules

Optical telescopes – biggest is best?

Sir— Twice in recent months, *Nature* has carried articles discussing future astronomical facilities^{1,2} but, although both articles mentioned 'value for money' (one in its title), no cost-benefit analysis was mentioned.

It seems strange that large sums of money are being spent on new astronomical facilities worldwide without any major cost-benefit analysis, covering both ground- and space-based observatories, to back up these investments. Abt has published a number of papers looking at some aspects of this, as have Trimble, and Martin and Irvine (see refs 3–5, for example), but it appears that no-one has yet undertaken a comprehensive survey. I am trying to rectify this by undertaking a cost-benefit analysis of both ground- and space-based observatories owned by governments and other institutions in the United States and British Commonwealth countries.

I have assessed the benefit or effectiveness of a particular facility by analysing the 15% most-cited papers published in the *Astrophysical Journal* and *Monthly Notices of the Royal Astronomical*

Society every half-year from 1958 to 1994 inclusive, at four-yearly intervals, and deducing which facilities were used to produce the new observations analysed in those papers. In many cases, more than one telescope or spacecraft were used to produce the data for a given paper and, in that case, a scoring system was used⁶ to take into account these different facilities, the total score for each paper being unity.

The annual costs dominate the capital costs for ground-based optical/infrared observatories, and Abt found³ that these annual costs vary as the telescope aperture to the power of 2.1. Using this synthetic, I have calculated the relative cost-effectiveness scores for different size categories of ground-based telescopes as $N_p/\Sigma d^{2.1}$ where N_p is the number of papers produced using data from telescopes in a given size category and $\Sigma d^{2.1}$ is the sum of the diameters to the power 2.1 of all the telescopes available in that size category.

The values of $N_p/\Sigma d^{2.1}$ have been calculated for each half-year and have then been normalized so that they total 100 in each half-year considered, to give each half-year an equal weighting.

The average of these relative cost-effectiveness scores for the second half of my period, 1978–94, was for ground-based telescopes in the above geographical area:

Telescope aperture range (in m)	Relative cost-effectiveness (arbitrary units)
2.55–5.08 (200")	45 ± 6
1.23–2.54 (100")	33 ± 7
0.62–1.22 (48")	17 ± 6
0.61 (24")	5 +8/–5

These results clearly show that 'big is best' when looking at cutting-edge research. That is not to say that there is no role for small telescopes, but at a national level we need to build and equip the largest optical/infrared telescopes we can afford.

David Leverington

Physics Department,
Open University,
Walton Hall,
Milton Keynes, MK7 6AA.
e-mail: D.Leverington@open.ac.uk

1. Reichhardt, T. et al. *Nature* **381**, 461 (1996).
2. *Nature* **383**, 651 (1996).
3. Abt, H.A. *Publ. astr. Soc. Pac.* **92**, 249 (1980).
4. Trimble, V. *Publ. astr. Soc. Pac.* **107**, 977 (1995).
5. Martin, B.R. & Irvine, J. *Res. Pol.* **12**, 61 (1983).
6. Leverington, D. Q. *J. R. astr. Soc.* **37**, 643 (1996).

Unequivocal spelling for brimstone

Sir— On the spelling of the sixteenth element of the periodic table, Morris recently drew attention to the Sanskrit 'sulvere' for sulphur (Nature **384**, 103; 1996). While this is intriguing, as an etymological basis for modern English spelling it is a red herring.

A historical account of the development of the two spellings is given by C. A. Michie and D. R. Langslow (*Br. Med. J.* **297**, 1697–1699; 1988), who show that, since the early Middle Ages, only English among European languages has used the Greek-derived -ph- rather than the -f- spelling. On the basis of one vote per language, therefore, the appropriate spelling for the European Union would be 'sulfur'. The relatively recent adoption of -f- in American English would be consistent with this.

We suggest, however, a more radical solution to discrepancies in spelling, which is to (re-)adopt the name 'brimstone', abbreviated naturally to Bs. In popular medicine, 'brimstone' was used well into this century, at least in the United Kingdom. In chemistry, elemental S would then resume its ancient name of 'brimstone', sulfuric acid would become 'brimstonic acid' (H₂BsO₄) and

concern for acid rain would recall biblical equivalents (Genesis: 19, 24 — "The Lord rained upon Sodom and upon Gomorrah brimstone and fire...") with frightening vividness.

David F. Cram

Centre for Linguistics and Philology,
University of Oxford,
Oxford, UK

W. John Cram*

Department of Biological
and Nutritional Sciences,
University of Newcastle upon Tyne.
Newcastle upon Tyne NE1 7RU, UK
e-mail: w.j.cram@ncl.ac.uk

*To whom correspondence should be addressed.

A fortune awaits

Sir— I commented some time ago (Nature **378**, 532; 1995) on the use of neem tree products as insecticides. Indians were at that time naturally objecting to foreign companies exploiting an indigenous Indian resource for commercial ends. The same is now happening with a yellow paste from the root of the turmeric plant, which has been used as a healing and antibiotic substance in India from time immemorial.

But those who are raising objections to foreign pharmaceutical companies exploiting turmeric for medical

applications overlook its use for an entirely different purpose forgotten by Indian urban middle classes but very much alive in village India, at least here in the south. Village women assume a deep yellow over whatever parts of their skin can be seen, because the periodical application of turmeric paste keeps them free of superfluous hair. It is also an effective fungicide, fungus infections of the skin being common in tropical climates.

Unfortunately, this is a messy business, liable to stain not only skin but washbasins and bathtubs, sanitary ceramics that are unknown in the villages. The practice has therefore been discontinued and has now probably been forgotten among affluent middle-class city dwellers.

A fortune awaits the pharmacist who can isolate the active depilative in turmeric.

Harry Miller

3A Satyanarayana Avenue,
Boat Club Road,
Madras, Tamil Nadu,
India 600 028
e-mail: mdsaab80@giasmd01.vsnl.net.in

corres@nature.com Letters submitted for Correspondence may be e-mailed to corres@nature.com. Do not send items intended for any other section by e-mail unless requested to do so.

Predicting the CJD epidemic in humans

Fourteen cases of new-variant Creutzfeldt–Jakob disease have so far been confirmed in the United Kingdom. Are they the start of an epidemic? If so, how informative will cases in the next few years be in predicting its course?

**S. N. Cousens, E. Vynnycky,
M. Zeidler, R. G. Will and P. G. Smith**

Since the description of ten cases of a new variant of Creutzfeldt–Jakob disease (vCJD) in the United Kingdom in March 1996 (refs 1, 2), four more UK cases and one French case have been confirmed³. It remains unknown whether the 'new form' of CJD is due to human infection with the agent responsible for bovine spongiform encephalopathy (BSE) in cattle^{1,2}. But the possibility that many infected cows have entered the human food supply⁴ has fuelled fears that there might be many cases of vCJD in the future. Too little is known about prion diseases to make, at this stage, confident predictions about what might happen in humans. But the rate at which a disease appears early in an epidemic can give clues about its future size.

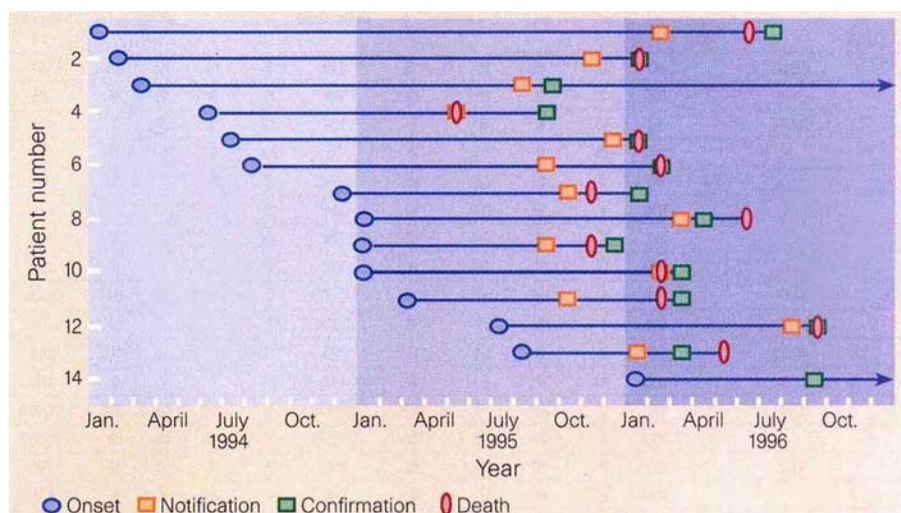
Assuming that vCJD is due to exposure to the BSE agent, we consider two related questions. Given that only 14 UK cases have been confirmed up to the end of 1996, with 13 showing clinical onset in 1994 or 1995, can we rule out the possibility of a large epidemic of vCJD? If not, how informative will cases in the next few years be in predicting the epidemic?

We used mathematical models that make assumptions about the distribution of exposure of the human population to the BSE agent over time and about the time from infection to clinical onset (incubation period) of vCJD. The models estimate, by back calculation⁵, how many human infections eventually resulting in disease would have been required to produce 13 cases of vCJD with onset in 1994/95 (see figure) and how many cases might arise in the next three years.

Table 1 Numbers of BSE cases

Year of clinical onset	No. of confirmed cases	No. of cases using maximum bounds on under-reporting
	(EPA)	(EPB)
1983	0	25
1984	0	50
1985	15	240
1986	63	1,052
1987	662	3,310
1988	3,238	4,760
1989	7,823	7,823
1990	14,693	14,693
1991	26,160	26,160
1992	37,488	37,488
1993	34,089	34,089
1994	22,957	22,957
1995	13,766	13,766
1996*	2,262	2,262

EP, exposure pattern. *Up to 31 March.



Dates of onset, death, referral and confirmation of diagnosis for 14 cases of new-variant CJD in the United Kingdom. For the case with most recent onset, referral and confirmation were almost simultaneous.

We assume that, until 1989, the number of people newly infected with the BSE agent each year was proportional to the number of cases of BSE with onset in that year (Table 1). This assumption is reasonable if bovine material is infectious to humans for only a short period before cattle develop clinical BSE. The order requiring the destruction of cattle clinically suspected of having BSE was not introduced until August 1988, and, until then, clinically affected animals may themselves have entered the human food chain. Two sets of figures for the number of BSE cases are used. The first set, the number of confirmed cases of BSE, assumes that the degree of under-reporting of BSE changed little over time (exposure pattern A). The second set (exposure pattern B) assumes that under-reporting was greatest early in the epidemic (maximum estimates, kindly provided by R. Anderson and N. Ferguson, based on their BSE epidemic model⁶).

In November 1989, UK legislation was introduced that banned specified bovine offals (SBO) from human consumption. We further assume that from 1990 the number of people newly infected each year, while remaining proportional to annual numbers of BSE cases, decreased by 90 or 100 per cent.

We assume that individuals infected with the BSE agent develop vCJD after a long and variable incubation period. Just how long and variable, we do not at present know. For many infectious diseases, the distribution of incubation periods is roughly lognormal⁶. Other distributions used to model incubation periods include the gamma and the Weibull^{4,7}. We used all three, unlagged, in our calculations.

The estimated number of infections is extremely sensitive to the shape of the incubation

period distribution, its mean and its variability; the number ranges from less than a hundred to many thousands (Table 2). The largest estimates arise with a long mean incubation period and little variation about the mean. Some important differences in the distribution curves greatly affect estimates of the epidemic size based on cases appearing early on. The lognormal distribution generally produces larger estimates than the gamma. The Weibull distribution produces an earlier rise in cases than the lognormal and gamma distributions and so implies a smaller total number of infections. Using the Weibull, the estimated number exceeds 1,000 in only a few situations, the largest being around 3,000.

Few data are available to guide our choice of incubation period for vCJD. But no cases of vCJD with onset before 1994 have been identified. If our model is to predict fewer than, say, five cases with onset before 1994, then the unlagged Weibull would not be a good approximation to the incubation period distribution and the mean incubation period of vCJD would be unlikely to be as short as five years. Several scenarios based on longer mean incubation periods would also be judged incompatible with no cases before 1994 (Table 2). But this restriction alone does not greatly reduce the range of possible outcomes.

For kuru, incubation periods of up to 30 years have been recorded⁸. The mean incubation period for CJD caused by contaminated human pituitary hormone is currently about 13 years⁹, although this may be an underestimate as cases with longer incubation periods may yet appear. The range of 'spreads' that we consider (90% of individuals developing disease within 1.5–2 times the mean incubation

Table 2 Predicted numbers of new-variant CJD cases under various possible scenarios

Mean incubation period (yr)	Ninetieth centile of incubation period distribution (yr)	Effectiveness of ban on specified bovine offals (%)	Lognormal incubation period distribution										Gamma incubation period distribution					
			Exposure pattern A					Exposure pattern B					Exposure pattern A			Exposure pattern B		
			Predicted cases with onset pre-1994	in 1996	in 1997	in 1998	Total no. of human infections	Predicted cases with onset pre-1994	in 1996	in 1997	in 1998	Total no. of human infections	Predicted cases with onset pre-1994	in 1996	in 1997	in 1998	Total no. of human infections	Total no. of human infections
10	15.0	90	3	12	15	18	213	5	10	12	13	151	4	11	14	17	211	156
			3	11	12	12	99	5	9	10	9	83	4	10	11	11	103	88
		100	3	8	8	7	104	17	7	7	7	101	15	7	8	7	117	114
			19	6	5	4	75	23	5	5	4	79	22	6	6	5	92	94
15	22.5	90	1	19	32	48	1,595	3	15	22	30	801	3	14	22	30	1,001	620
			1	18	29	40	714	3	14	21	27	430	3	13	19	24	469	342
		100	10	8	9	10	174	13	8	8	9	158	13	8	8	9	188	177
			13	7	6	6	107	16	6	6	6	107	19	7	7	6	137	137
20	30.0	90	1	26	54	97	12,000	2	19	34	54	5,000	2	16	27	41	4,000	2,179
			1	25	51	89	5,000	2	19	33	51	2,421	2	15	24	35	1,800	1,189
		100	8	9	10	11	284	11	8	9	10	244	13	8	9	9	276	255
			10	7	8	8	162	13	7	7	7	156	18	7	7	7	195	191
25	37.5	90	1	32	79	166	80,000	2	22	45	83	24,000	2	18	31	50	13,000	7,000
			1	31	76	156	35,000	2	22	45	80	13,000	2	17	28	44	6,000	4,000
		100	7	10	11	13	450	9	9	10	11	370	12	8	9	10	380	346
			8	8	9	9	245	11	8	8	8	228	17	7	7	8	263	256

Numbers greater than 2,500 have been rounded to the nearest 1,000. Numbers greater than 25,000 have been rounded to the nearest 5,000.

period) corresponds closely with the range of 'dispersion factors' reported for a variety of acute infectious diseases⁶ and is consistent with that used to model the BSE epidemic⁴. A mean incubation period for vCJD of around 15 years points to an epidemic of between several hundred and several thousand cases. But a longer mean incubation period or a 'narrower' incubation period distribution allows for a much larger epidemic.

If under-reporting of BSE cases was common at the start of the epidemic and then declined (exposure pattern B), then we would be further into any BSE-related vCJD 'epidemic' than if under-reporting remained constant over time (exposure pattern A). Exposure pattern B therefore generally leads to lower estimates of the total number of infections. If, say, the distribution of vCJD incubation periods is lognormal, with a mean of 10 years and 90 per cent of infected individuals developing the disease within 15 years, and the SBO ban was 90% effective, exposure pattern B would indicate a total epidemic of about 151 infections (Table 2). Exposure pattern A would indicate a total epidemic of about 213 infections.

If bovine material becomes infectious to humans well before clinical onset of BSE, humans would have been exposed to the BSE agent earlier, relatively, than we have assumed. This would put us further into the vCJD 'epidemic' and so would generally lead to lower estimates of human infections. Although we assume that the SBO ban was either 90% or 100% effective, our results do not critically depend on this assumption. If, say, the SBO ban was only 50% effective, the estimated number of infections in the lognormal model is two to three times greater in the scenarios examined in Table 2. This difference is small compared with the effect of varying

the assumptions about incubation period.

If we allow for variability in onset times or the identification of further cases with onset in 1994/95, an even wider range of outcomes is possible. The 13 cases identified are compatible with an underlying expected number of cases with onset in 1994/95 of anything from 6 to 20 (95% confidence interval), resulting in a proportional decrease or increase in the numbers in Table 2. Given the typically long delay between onset and confirmation (see figure), further cases with onset in 1994/95 are likely. If cases continue to accrue at the present rate, the final number with onset in 1994/95 might be about 23 (C. P. Farrington and D. De Angelis, personal communication).

We hesitate to draw sweeping conclusions from calculations based on few data and several currently unverifiable assumptions. Enormous uncertainty inevitably surrounds any modelling when only 14 cases of the disease have been confirmed and without good information about the incubation period distribution. Still, if cases of vCJD are due to exposure to the BSE agent, as recent evidence suggests¹⁰, two tentative conclusions may be drawn. First, it would be premature to conclude that because only 14 UK cases have been confirmed so far, any subsequent epidemic will necessarily be small. Second, although the numbers of cases over the next few years may provide a better indication of how large any epidemic might eventually be, much uncertainty may remain even in four years' time.

If, for example, the number of vCJD cases with onset in each of the next three years is roughly constant and less than 20 a year, the final size of the epidemic may well be a few hundred cases or less. Alternatively, if there are 25 or more cases with onset in 1996, with a

doubling or tripling in each of the following years, this would be compatible with a long mean incubation period and an epidemic of many thousand cases. On the other hand, 20 or so cases in 1996, 30–35 in 1997 and about 50 in 1998 would be compatible with both a lognormally distributed incubation period with a mean of 15 years and a ninetieth centile at 22.5 years (around 1,600 infections) and a gamma-distributed incubation period with a mean of 25 years and a ninetieth centile at 37.5 years (around 13,000 infections) (exposure pattern A; SBO ban, 90% effective).

If it is shown that vCJD is due to exposure to the BSE agent, then improved estimates of the incubation period distributions of the two closest analogies we have to vCJD – kuru and CJD in recipients of pituitary hormones – will become a high priority. So too will estimates of the routes and the amount of infectious material entering the human food supply. □

S. N. Cousens, E. Vynnycky and P. G. Smith are in the Department of Epidemiology and Population Sciences, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK.

M. Zeidler and R. G. Will are at the National CJD Surveillance Unit, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK.

Acknowledgement: We thank R. Kimberlin and P. Fine for comments on earlier versions of the manuscript.

Correspondence should be addressed to S. N. C.

1. Statement by the UK Spongiform Encephalopathy Advisory Committee (House of Commons, 20 March 1996; HMSO).

2. Will, R. G. *et al. Lancet* 347, 921–925 (1996).

3. Chazot, G. *et al. Lancet* 347, 1181 (1996).

4. Anderson, R. M. *et al. Nature* 382, 779–788 (1996).

5. Brookmeyer, R. *Biometrics* 52, 781–796 (1996).

6. Sartwell, P. E. *Am. J. Hyg.* 51, 310–318 (1950).

7. Marion, S. A. & Schechter, M. T. *Stat. Med.* 12, 617–631 (1993).

8. Klitzman, R. L., Alpers, M. P. & Gajdusek, D. C. *Neuroepidemiology* 3, 3–20 (1984).

9. Brown, P., Preece, M. A. & Will, R. G. *Lancet* 340, 24–27 (1992).

10. Collinge, J., Sidle, K. C. L., Meads, J., Ironside, J. & Hill, A. F. *Nature* 383, 685–690 (1996).

Cope's rule as psychological artefact

Stephen Jay Gould

According to a 'law' formulated by E. D. Cope in 1871, the body size of organisms in a particular evolutionary lineage tends to increase. But Cope's rule has failed the most comprehensive test yet applied to it.

Two strategies have dominated our approach to discovering generalities about the history of life at broadest palaeontological scales. We have, in a first and more conventional manner, developed putative laws and theories by the usual route of elucidating the universal properties of matter and motion, and then testing these principles (such as natural selection) by their ability to produce the phenomena of life's history. The second, almost opposite, strategy tries to induce generalities of high relative frequency directly from recurrent phenomena of the fossil record itself.

This second strategy, once highly popular, led to a panoply of propositions well known to past generations of scholars (Williston's law of the reduction of serially repeated structures to fewer sets of differentiated organs; Dollo's law of the irreversibility of evolution). But these laws are now largely forgotten, either because they have been proven untrue, or because they have been better viewed as a local consequence of a broader theory discovered by the first strategy.

However, one 'law' of this older tradition has survived to become an important generality about life's history: Cope's rule of phyletic size increase within lineages. This principle supposedly applies to all groups of organisms¹: unicells and multicells, invertebrates and vertebrates, the marine and terrestrial, and the cold- and warm-blooded. Not only has the factual validity of this rule been widely accepted, but traditional literature knows only one mode of potential explanation: if body size tends to increase, then large-bodied individuals must, *ceteris paribus*, be generally favoured by natural selection. Hallam, for example, recently wrote²: "since phyletic size increase is such a widespread trend in the animal kingdom, there must be manifestly one or more selective advantages of larger size.... Among those proposed are an improved ability to capture prey or ward off predators; greater reproductive success; increased regulation of the internal environment; and increased heat regulation per unit volume".

But a generation of palaeobiological

study has fruitfully focused on exposing the many things in Heaven and Earth undreamed of in strict darwinian philosophy — particularly the numerous, shaping phenomena of life's history that cannot be explained as simple extrapolations, through immense time, of gradual adaptive transformation within individual populations (mass extinction, punctuated equilibrium, and so on). Such a tradition must necessarily call into question the putative generalities that have always been so explained, for example Cope's rule as rendered by general adaptive advantages of larger size for individual organisms.

One mode of challenge, first proposed by Stanley³ in 1973, accepts the phenomenon but develops a radically different explanation based on constraints within groups of species rather than accumulated adaptive propensities of individual lineages. If founders of large monophyletic lineages tend to be small-bodied and near some lower limit for the basic body plan, as Cope himself proposed in another 'law of the unspecialized'⁴, then a general increase in mean body size among species of a clade may arise as a drift away from a lower limit, rather than a directed trend towards an upper bound. Indeed, in such cases, as for planktonic Foraminifera (data of R. Norris reported in ref. 5), the modal size among species may never depart from the initial size of the common ancestor, even while the mean size among species increases.

But we should also ask a different, even more basic, Emperor's-new-clothes question, as Jablonski does in the paper on page 250 of this issue⁶: is Cope's rule even true at all? One would think that issues so fundamental, and so eminently testable, had been conclusively resolved long ago — except for a perverse trait of the human psyche. We tend to pick most 'notable' cases out of general pools, often for idiosyncratic reasons that can only distort a proper scientific investigation.

For example, don't we all have the impression that our own birthdate appears at an uncommonly high relative frequency

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Law man — Edward Drinker Cope of the eponymous rule.

among cited dates for events in news stories — surely an effect of what we notice and what we pass by in a random array? Similarly, might not our conviction about the validity of Cope's rule be a psychological artefact of singling out lineages that display size increase because we all know that 'bigger is better'? Such a procedure might provide an example of another pervasive and lamentable bias of human reasoning: our tendency to focus on extremes that intrigue us, rather than full ranges of variation⁷. The obvious test requires that we abandon our habit of selective search for the expected and, instead, study all lineages in large clades with excellent data over substantial geological intervals. Jablonski has followed this admirable procedure in the most comprehensive set of data ever assembled to test Cope's rule — and the rule fails in this case.

Jablonski studied all 191 bivalve and gastropod lineages of molluscs with sufficient data (including 1,086 species) during 16 million years of Late Cretaceous time for the rich faunas of the Gulf and Atlantic coastal plain of North America. He compiled the range of adult sizes at the beginning and end of this interval across all species within each lineage (with the size for each species represented by the geometric mean of height and length for the largest specimen in each species). He consistently concentrated on complete ranges of variation, rather than expansion or contraction of extreme values alone.

Jablonski found, first of all, that lineages showing net increase in size (that is, increase in both the smallest and largest species) are no more common than lineages showing net decrease of both smallest and largest: 27–30% display net increase; 26–27% display net decrease. Moreover, in the additional 25–28% of lineages that show size increase for the largest species, the smallest species also decrease in size over the same interval — yielding a pattern of expansion in the overall range of variation (at both high and low ends), not a directed trend towards increas-

ing general size. In short, although many individual lineages do show increase in body size, just as many decrease. So a full account of all data provides no support for Cope's rule as a preferential bias in the evolution of size.

The general issue thus addressed is neither quirky nor insignificant. Our strong and biased predilection for focusing on extremes (and misconstruing their trends as surrogates for a totality), rather than documenting full ranges of variation, generates all manner of deep and stubborn errors. Most notable of these misconceptions is the false and self-serving notion that evolution displays a central and general thrust towards increasing complexity, when life, in fact, has been dominated by its persistent bacterial mode for all 3.5 billion years of its history on

Earth. We should remember Little Buttercup's admonition to Captain Corcoran in *H.M.S. Pinafore*, that "things are seldom what they seem", while we must shun the allure of bigness, for "bulls are but inflated frogs".

Stephen Jay Gould is in the Museum of Comparative Zoology, The Agassiz Museum, Harvard University, 26 Oxford Street, Cambridge, Massachusetts 02138, USA.

1. Newell, N. D. *Evolution* 3, 103–124 (1949).
2. Hallam, A. in *Causes of Evolution* (eds Ross, R. M. & Allmon, W. D.) 249–269 (Univ. Chicago Press, 1990).
3. Stanley, S. M. *Evolution* 27, 1–26 (1973).
4. Cope, E. D. *The Primary Factors of Organic Evolution* (Open Court, Chicago, 1896).
5. Gould, S. J. *J. Paleontol.* 62, 319–329 (1988).
6. Jablonski, D. *Nature* 385, 250–252 (1997).
7. Gould, S. J. *Full House: The Spread of Excellence from Plato to Darwin* (Harmony, New York, 1996).

Creutzfeldt–Jakob disease

Epidemic or false alarm?

David C. G. Skegg

By a quirk of fate, it was during Britain's National Science Week in 1996 that the Secretary of State for Health made an announcement that focused more public attention on science than any other event in this decade. The disclosure that ten cases of a new variant of Creutzfeldt–Jakob disease (vCJD) might be related to the epidemic of bovine spongiform encephalopathy (BSE) prompted fears of a disaster of apocalyptic proportions. The scientific publication that followed¹ produced an element of anticlimax, because the suggestion of a causal link rested mainly on coincidences — of time and geography — as well as on biological plausibility, which was still highly contentious.

Fourteen cases of vCJD have now been confirmed in Britain. If one swallow does not make a summer, do 14 cases constitute an epidemic? On page 197 of this issue Cousens *et al.*² present a simple epidemic model of the new variant, based on the assumption that vCJD is caused by exposure to the BSE agent.

Although vCJD produces an unusual clinical picture, its most distinctive feature is the brain pathology, which involves florid plaques — large deposits of abnormal prion protein (PrP) surrounded by a zone of vacuolated tissue. When macaque monkeys are intracerebrally inoculated with brain homogenate from cattle with BSE, they develop a disease that has a similar pathology to vCJD³. Moreover, molecular analysis has shown that vCJD has strain characteristics that are distinct from other types of CJD, but similar to those of BSE that has been transmitted to the mouse, to the domestic cat and to the macaque⁴. The biological underpinning of the hypothesis about vCJD is thus stronger than it was a year ago.

Although less attention has so far been given to epidemiological research, Will and Zeidler⁵ argue that this may hold the key to understanding any link between BSE and human disease.

It may seem strange to attempt any modelling study on the basis of only 14 cases, especially when we know nothing about the distribution of incubation periods — the shape of the distribution curve, its mean (in years), or its spread. Yet scientists, as well as politicians, have been tempted to draw comforting conclusions from the fact that only 14 cases have been identified in Britain. Cousens *et al.*² show that, with plausible assumptions about the incubation period and patterns of exposure, the total number of cases that can be predicted ranges from about a hundred to tens of thousands. Perhaps more surprising is that, even in another four years' time, considerable uncertainty may remain.

From an epidemiological viewpoint, the evidence linking vCJD with BSE is weak because it is based only on the temporal and geographical association between the two diseases in whole populations. One would wish to see analytical studies showing a higher risk of developing vCJD in individuals who had been more exposed to the BSE agent. Apart from the small number of cases available for study, such research is difficult when the relevant exposures are uncertain. The media have published colourful anecdotes about the fondness of certain patients for hamburgers and other fast foods, but such information is clearly subject to bias and we have no comparable data about others of the same age. Cousens *et al.* give no information about diet, nor do they state whether the most recent patients resembled earlier ones in being homozygous for

methionine at codon 129 of the PrP gene^{1,6}.

It would be a failure of perspective to focus on the epidemiology of CJD in Britain alone. When a 26-year-old French man was diagnosed as having vCJD, the possible causal relationship with BSE was called into question⁶. This argument was surprising because, apart from the existence of BSE in France (with some controversy about the extent of under-reporting), France has imported beef and live calves from Britain.

In fact, isolated cases of the new variant, even in countries without BSE (or in the years before the onset of the BSE epidemic), would not invalidate the causal hypothesis. One theory about the origin of BSE is that it was a rare disease in cattle that came to attention only after its prevalence was amplified by modern feeding practices⁷. Yet not a single case of vCJD was found in the United States last year, despite intensified surveillance⁸. There are excellent diagnostic facilities in the US, which has a population that is over four times larger than that of Britain — so 14 cases over two years in Britain would be equivalent to about 60 cases in the US. The fact that no cases of vCJD have emerged in a country that does not have BSE is one of the most significant pieces of the epidemiological jigsaw.

Neuropathologists in many countries have been re-examining their archival material, yet there are no new reports of previously unsuspected cases of vCJD. This adds weight to the conclusion that a distinct, new entity has emerged in Britain and France, and it no longer seems credible that the recognition of vCJD might simply reflect improved detection⁹. Despite the preliminary nature of the epidemiology, the recent biochemical evidence⁴, combined with a lack of plausible alternatives, makes exposure to BSE a likely explanation.

Albert Camus wrote in *The Plague* that "nothing is less sensational than pestilence and by reason of their very duration great misfortunes are monotonous". The progress of an epidemic is especially slow when the incubation period is measured not in days, but in years or decades. There is still room for optimism that vCJD may remain a very rare disease, but Cousens *et al.*² remind us that it is still too early to tell.

David C. G. Skegg is in the Department of Preventive and Social Medicine, University of Otago, PO Box 913, Dunedin, New Zealand.

1. Will, R. G. *et al.* *Lancet* 347, 921–925 (1996).
2. Cousens, S. N., Vynnycky, E., Zeidler, M., Will, R. G. & Smith, P. G. *Nature* 385, 197–198 (1997).
3. Lasmézas, G. I. *et al.* *Nature* 381, 743–744 (1996).
4. Collinge, J., Sidle, K. C. L., Meads, J., Ironside, J. & Hill, A. F. *Nature* 383, 685–690 (1996).
5. Will, R. & Zeidler, M. *Br. Med. J.* 313, 833–834 (1996).
6. Chazot, G. *et al.* *Lancet* 347, 1181 (1996).
7. Kimberlin, R. H. *Rev. Sci. Tech. Off. Int. Epiz.* 11, 347–390 (1992).
8. Holman, R. C., Khan, A. S., Belay, E. D. & Schonberger, L. B. *Emerging Infect. Dis.* 2, 333–337 (1996).
9. Roberts, G. W. & James, S. *Curr. Biol.* 6, 1247–1249 (1996).

Life on other moons

Christopher F. Chyba

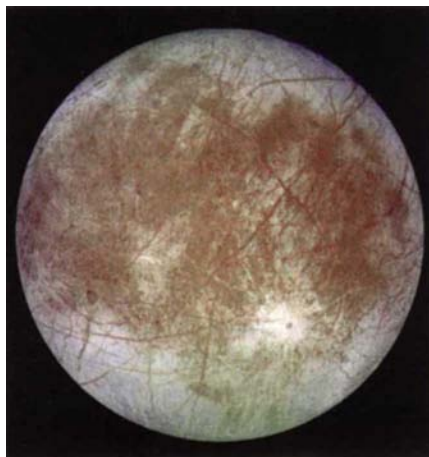
What does it take for a world to be habitable? As with so many other questions in exobiology, Carl Sagan helped pioneer this topic¹, and in an article last year² he presented a masterly and cautionary review of the changing scientific fashions on the question. Now that extrasolar giant planets (or brown dwarfs) have been found around nine main-sequence stars, Williams *et al.*, on page 234 of this issue³, extend the discussion to the habitability of moons around these worlds.

What it takes for a world to be habitable depends on who will be doing the inhabiting. Clearly, the requirements of humans differ from those of forests⁴, which in turn are more stringent than those of, say, green slime. Habitability for putative extraterrestrial technical civilizations may be the most interesting kind to consider. Yet since the history of surface life on Earth is largely the history of green slime, its needs must remain of special interest to exobiologists.

Recent work has defined 'habitable zone' as the range of orbital distances within which worlds can maintain liquid water on their surfaces^{5,6}. This reflects the utter dependence of terrestrial life on liquid water⁷. The giant planets orbiting the stars 16 Cyg B and 47 UMa may by this definition lie within their stars' habitable zones. But Williams *et al.* argue that these worlds (like the gas giants in our own system, whose solid cores lie at depths where pressures and temperatures are extreme) probably lack a solid or liquid surface suitable for life.

So Williams *et al.* instead consider the possible moons of the extrasolar giants, and show that some of them might provide habitats for extraterrestrial life. They find that such a moon would need to be a bit more massive than Mars (about 0.1 Earth masses) in order to retain a substantial atmosphere (essential for surface liquid water) for billions of years. The moon might also need an appreciable magnetic field to protect its atmosphere against loss due to sputtering by charged particles. Jupiter's moon Ganymede, the largest moon in our Solar System, has a magnetic field, but is only 0.03 of Earth's mass. More massive moons may well exist.

There is a certain irony in postulating moons around the newly discovered planets as sites for life. Before the recent flurry of evidence for planetary systems, the argument for extrasolar habitable planets ran like this: our star appears to be typical; it has planets suitable for life; so other stars probably do as well. We now know that other stars have



Europa, as seen by the Galileo spacecraft. Europa, Jupiter's second moon, is one of the most promising sites for life in the Solar System, for despite having a frozen surface there is evidence of an ocean below. But moons around the giant planets in other solar systems may be more like Earth, with surface liquid water and moderately thick atmospheres.

planets, but current observing techniques cannot discover companions much below the mass of Jupiter, which is widely deemed unsuitable for life. We therefore now have an argument for habitable moons that parallels the previous one for planets: our giant planets appear to be typical; they have moons that might be suitable for life (were they more massive and within the Sun's habitable zone, by the arguments of Williams *et al.*); so extrasolar giant planets (in habitable zones) probably do as well. It is certainly of the first importance for exobiology that we now have evidence that planetary formation is common. Nevertheless, ultimately what is needed is a means of detecting Earth-sized worlds — a capability that is years, and possibly decades, away⁸.

It should be emphasized just how conservative the definition of habitable zone used by Williams *et al.*³ (and most other workers) really is. Carl Sagan was famous for his concern that our notions of the conditions needed for extraterrestrial life are cautious to the point of being chauvinistic. In his recent piece on habitable zones², he raised a series of objections to the 'surface liquid water' definition of habitability.

For example, Sagan and Salpeter have speculated on possible ecological niches on giant planets: perhaps water clouds, simple organic molecules and abundant energy sources are enough for life, and no solid or liquid surface is required⁹. To this, the counter-objection has been raised that no

airborne ecologies are known on Earth: if such ecologies are possible, why are there apparently no terrestrial clouds green with microorganisms¹⁰?

Less speculatively, it is known that deep subsurface microbial ecologies exist on Earth. If Earth is in fact home to a 'deep, hot biosphere'¹¹ not dependent on solar energy, such biospheres might exist on other worlds well outside the surface-liquid-water zone — unless surface liquid water or surface energy sources (principally solar ultraviolet) are required for the origin of life.

In our own Solar System, circumstantial evidence for a liquid subsurface ocean on Jupiter's moon Europa is growing, and there could be regions of Europa where conditions lie within the range of adaptation of Antarctic terrestrial organisms¹². Although such speculations will remain unresolved until the question of a European ocean is settled, it is striking that Europa, one of two prime candidates for a second habitable world in our own Solar System, nevertheless lies well beyond the surface-liquid-water zone, has only about a tenth of Mars's mass, and has almost no atmosphere.

Our ignorance is sufficient to allow even more speculative environments within our Solar System¹³. As Sagan pointed out, if even a conservative definition of habitable zone suggests many locales for life around other stars — a conclusion to which Williams *et al.* lend support — then "whatever we have ignored could only serve to broaden the biological arena"².

As it stands, Earth is the only known habitable world. Investigating the conditions required for worlds to be habitable, naturally reminds us of our responsibility to our own. Here, too, Carl Sagan's voice will be dearly missed.

Christopher F. Chyba is in the Department of Planetary Sciences, University of Arizona, Tucson, Arizona 85721, USA.

1. Shklovskii, I. S. & Sagan, C. *Intelligent Life in the Universe* (Holden-Day, San Francisco, 1966).
2. Sagan, C. in *Circumstellar Habitable Zones* (ed. Doyle, L. R.) 3–14 (Travis House, Menlo Park, CA, 1996).
3. Williams, D. M., Kasting, J. F. & Wade, R. A. *Nature* **385**, 234–236 (1997).
4. Heath, M. in *Circumstellar Habitable Zones* (ed. Doyle, L. R.) 445–457 (Travis House, Menlo Park, CA, 1996).
5. Kasting, J. F., Whitmire, D. P. & Reynolds, R. T. *Icarus* **101**, 108–128 (1993).
6. Kasting, J. F. in *Circumstellar Habitable Zones* (ed. Doyle, L. R.) 17–28 (Travis House, Menlo Park, CA, 1996).
7. McKay, C. P. *Icarus* **91**, 93–100 (1991).
8. Burke, B. F. (ed.) *TOPS: Toward Other Planetary Systems* (NASA, 1992).
9. Sagan, C. & Salpeter, E. E. *Astrophys. J. Suppl. Ser.* **32**, 737–755 (1976).
10. McKay, C. P. in *Van Nostrand Reinhold Encyclopedia of Planetary Sciences and Astrogeology* (eds Shirley, J. H. & Fairbridge, R. W.) (Van Nostrand Reinhold, New York, in the press).
11. Gold, T. *Proc. Natl. Acad. Sci. USA* **89**, 6045–6049 (1992).
12. Reynolds, R. T., Squyres, S. W., Colburn, D. S. & McKay, C. P. *Icarus* **56**, 246–254 (1983).
13. Chyba, C. F. & McDonald, G. D. *Annu. Rev. Earth Planet. Sci.* **32**, 215–249 (1995).

Molecular fish on chips

André Goffeau

More than 100,000 new gene sequences — from bacteria to humans — should become available in the decade spanning 1995–2005. Two complete small bacterial genome sequences, *Haemophilus influenzae* (1.7 Mbp) and *Mycoplasma genitalium* (0.6 Mbp); the *Saccharomyces cerevisiae* genome (12 Mbp); and the sequences of the archaeobacterium *Methanococcus jannaschii* (1.8 Mbp) and the blue-green bacterium *Synechococcus* (3.6 Mbp) genomes are already available. Many other sequences should be completed in 1997 and 1998, including the long-awaited *Escherichia coli* (4.7 Mbp) and *Bacillus subtilis* (4.2 Mbp) genomes, and that of the worm *Caenorhabditis elegans* (about 100 Mbp). Next in line are the genomes of the plant *Arabidopsis thaliana* (100–150 Mbp) and the fruitfly *Drosophila melanogaster* (100–150 Mbp), and the complete *Homo sapiens* genome (~3,000 Mbp and 100,000 genes) is optimistically expected for the year 2005.

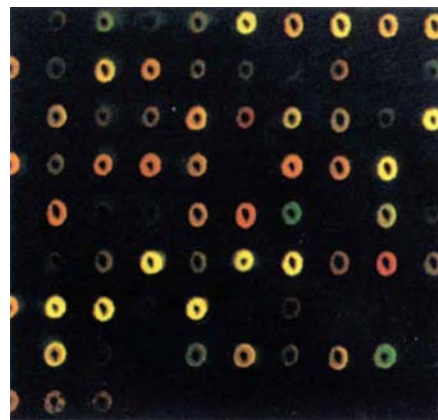
New methods are needed for the analysis of this unprecedented deluge of data, and a flood of papers^{1–8} now report the use of high-density oligonucleotide arrays for large-scale analysis of genomic data (see table). The principle is simple — small oligonucleotide probes, 15–20 bases in length, are synthesized and bound in ordered arrays to glass or nylon supports. The probes serve as bait for molecular fishing, by hybridizing to fluorescently labelled complementary DNAs called targets. The pattern and extent of the fluorescence that is observed gives information on the nucleotide sequence of the targets, as well as on the amounts of a particular cDNA that is present in the mixture of targets analysed.

Molecular hybridization of polynucleo-

tides — based on the spontaneous formation of hydrogen bonds between A and T or G and C nucleotides — has been used for more than 20 years and forms the basis of classic methods such as Sanger DNA sequencing and the polymerase chain reaction. The emerging applications for molecular hybridization have been improved by the latest technologies, such as combinatorial generation of polynucleotides; parallel phosphoramidite-based DNA synthesis on the solid support; photolithographic masking (the principle of which is borrowed from the microchip industry); and sensitive automated scanning of the fluorescence by confocal microscopy.

The new papers report variations on the model protocol described above. In one report published in *Nature Biotechnology*¹, 21 messenger RNA targets from the entire RNA population of a mouse T-cell line were quantitatively monitored using an array of 65,000 different 20-mer oligonucleotide probes that were designed to hybridize with a total of 114 known mouse genes. Changes in expression were detected for 20 other mRNAs after the induction of cytokine production. RNAs present at a frequency higher than 1:300,000 were detected, and the amounts of mRNA could be quantified up to a frequency of 1:300.

Another study, published in *Nature Genetics*², was aimed at analysing gene expression patterns in human cancer, and 870 different cDNA probes from a human melanoma cell line were printed robotically onto a glass microscope slide (see figure). By comparing the hybridization patterns obtained with a mixture of two cellular mRNA populations that were marked with different fluorors, gene expression was



Fluor power — close-up of a DNA microarray, as used by DeRisi and colleagues² in their studies of gene expression in a tumorigenic cell line and its non-tumorigenic derivative.

analysed in cells whose tumorigenic properties had been suppressed by the introduction of a normal human chromosome⁶.

The microhybridization techniques can also be used in screening for human disease-related mutations. Another report in *Nature Genetics*³ reveals how arrays of 96,000 oligonucleotides were used to screen for a total of 24 heterozygous mutations (or single nucleotide polymorphisms) in exon 11 of the hereditary breast and ovarian cancer gene *BRCA1*. The sensitivity and specificity of the method were improved by differential colour analysis of the reference and test sample signals.

The use of short oligonucleotide arrays to sequence DNA by hybridization with large DNA fragments of unknown sequence will remain a challenge for a long time. But comparative sequencing by hybridization is now possible. In a study by the Affymetrix company⁴, the entire human mitochondrial genome — 16.6 kbp long — was re-sequenced by hybridization to an array that contained ~135,000 oligonucleotides positioned 35 nm apart. Up to 99 per cent of the mitochondrial sequence could be read correctly. Moreover, a total of 505 polymorphisms of the mitochondrial DNA were detected from ten African subjects. So high-density oligonucleotide arrays can be used to characterize, in parallel, sequence variations of several genes in a population.

The most innovative use of the high-density oligonucleotide arrays is 'molecular bar-coding', which exploits two features that are so far unique to yeast — the availability of the complete genome sequence and the ability to replace genes with markers by homologous recombination. The technique could possibly be used to determine the biological functions of the many newly identified yeast open reading frames. In a pilot study published in *Nature Genetics*⁵, 11 yeast genes of known function were replaced by drug-resistance marker cassettes, each of which was labelled with a unique 20-mer

New developments in biochip technology

Target	Assay	Reference
mRNA from mouse lymphocytes	Expression levels after induction of cytokine production	Lockhart, D. J. <i>et al.</i> ¹
mRNA from human melanoma cells	Expression levels after suppression of tumorigenicity	DeRisi, J. <i>et al.</i> ²
DNA from human germ lines	Mutations in the breast cancer gene <i>BRCA1</i>	Hacia, J. G. <i>et al.</i> ³
Mitochondrial DNA from human tissues	Polymorphisms in individual patients	Chee, M. <i>et al.</i> ⁴
DNA from yeast auxotrophic deletants	Bar-coding of deleted mutants	Shoemaker, D. D. <i>et al.</i> ⁵
DNA from human cytokine genes	Improved fluorescence detection using optical fibres	Ferguson, J. A. <i>et al.</i> ⁶
Oligonucleotides	Improved chemistry for oligonucleotide purity	Weiler, J. & Hoheisel, J. D. ⁷
Oligonucleotides	Improved stability of DNA/DNA duplexes	Hoheisel, J. D. ⁸

oligonucleotide. Because the selected strains were all auxotrophic mutants, they were analysed under specific competitive growth conditions. To determine the survival of each strain, the bar-code tags were amplified by the polymerase chain reaction, fluorescently labelled and then hybridized to a high-density oligonucleotide array. This contained, at defined positions, 20-mer oligonucleotides that were complementary to the tag sequences, and survival was measured by quantification of the fluorescence.

The success of this pilot experiment means that, in theory, all 6,000 of the yeast genes could be individually deleted and survival could be monitored under a variety of stress-inducing conditions such as pH, temperature, salts, drugs, X-rays and ultraviolet light. Through such a project it should also be possible to construct and characterize double, or other multiple, yeast mutants.

With the present technology it should be possible to synthesize, *in situ* and in less than a day, 400,000 different 20-mer oligonucleotides on a 1.6-cm² glass slide — each spot being 20 mm apart from its neighbours. But the ultimate challenge is to synthesize (and read) on a single chip the 4 million probes that would correspond, in parallel, to all 100,000 human genes. Arrays with synthesis sites as little as 1 mm apart might be needed, so technological improvements are clearly necessary. For example, the nucleotide probes that are synthesized by photochemical methods⁹ contain impurities that reduce their specificities and signal-to-noise ratios. Although improvements in the phosphoramidite chemistry allow the synthesis of high-quality oligonucleotides⁷, these methods are not yet compatible with large-scale, light-directed synthesis of high-density arrays.

Another problem is that the light-driven phosphoramidite-based oligonucleotides form duplexes of heterogeneous stability. Stability depends on the base composition (that is, the G/C/A:T ratio), although another new study has shown that the DNA-binding stability of short oligonucleotides can be equalized by the controlled substitution of ribonucleotides into the deoxyribonucleotide arrays⁸. But this technique cannot be used for more complex probes such as long oligonucleotides or full mRNAs which are often unstable and so have limited sensitivity. Moreover, complex probes and targets may form secondary or tertiary structures that prevent or reduce duplex formation, and so give a false-negative signal¹⁰.

Other problems include the fact that the light-directed oligonucleotide synthesis needs sophisticated and expensive hardware — at the manufacturing level a lithography machine and an oligonucleotide synthesizer are needed and, at the user level, a sophisticated confocal microscope is required. More seriously, at present, spots that are located less than 7.5 mm apart cannot be resolved.

Several new methods have been tested for improved sensitivity, among them the use of optical fibres⁶. In one experiment, an oligonucleotide probe was covalently immobilized on one end of an optical fibre and a sensitive imaging system was connected at the other end. Seven fibres, each with a different immobilized probe on its tip, were bundled together to form a multiple cytokine DNA sensor. The method was fast (taking less than 10 min), reproducible and sensitive (able to detect as little as 10 nM DNA). However, the relatively large size of the optical fibre is a limitation compared with the much denser chip arrays.

Many companies are developing technical solutions to these problems. A chip for the diagnosis of HIV mutations is ready for release and soon a variety of biochip arrays for the diagnosis of several forms of cancer and of hereditary or infectious diseases will become available. Not only will the exact nucleotide defects of known genes be detect-

ed, but modifications in the expression of a battery of genes that are influenced by a given disease will be quantified. And indeed, it does not seem unrealistic to hope that in the long run it will become routine to simultaneously diagnose — by DNA hybridization on biochips — modifications in the entire genomes of individual patients.

André Goffeau is in the Unité de Biochimie Physiologique, Université Catholique de Louvain, Place Croix du Sud, 2/20, 1349 Louvain-la-Neuve, Belgium.

1. Lockhart, D. J. *et al. Nature Biotechnol.* **14**, 1675–1680 (1996).
2. DeRisi, J. *et al. Nature Genet.* **14**, 457–460 (1996).
3. Hacia, J. G., Brody, L. C., Chee, M. S., Fodor, S. P. A. & Collins, F. S. *Nature Genet.* **14**, 441–447 (1996).
4. Chee, M. *et al. Science* **274**, 610–614 (1996).
5. Shoemaker, D. D., Lashkari, D. A., Morris, D., Mittmann, M. & Davis, R. W. *Nature Genet.* **14**, 450–456 (1996).
6. Ferguson, J. A., Boles, T. C., Adams, C. P. & Walt, D. R. *Nature Biotechnol.* **14**, 1681–1684 (1996).
7. Weiler, J. & Hoheisel, J. D. *Anal. Biochem.* **243**, 218–227 (1996).
8. Hoheisel, J. D. *Nucleic Acids Res.* **24**, 430–432 (1996).
9. Pirrung, M. C. & Bradley, J. C. *J. Org. Chem.* **60**, 6270–6276 (1995).
10. Southern, E. M. *et al. Nucleic Acids Res.* **22**, 1368–1373 (1994).

Colloids

When like charges attract

Cherry A. Murray

Larsen and Grier, on page 230 of this issue¹, show that two similarly charged polymer spheres suspended in water can attract each other when they are several diameters apart. This surprising result casts some light on a tricky theoretical many-body problem that has been swept under the rug for a century, and it has implications for colloids in nature and in industrial processes.

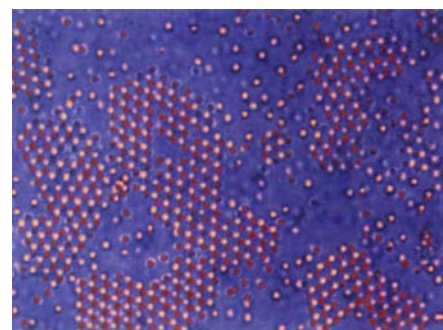
Colloidal suspensions of microscopic solid spheres can mimic the complex many-body properties of condensed matter on the atomic scale, particularly structural phase transitions². Interest in these colloids as model systems has increased in the past decade or so, because of the ease of obtaining highly monodisperse solutions (with diameter variation less than one per cent) and the use of sophisticated digital imaging.

Most colloids — the environment inside cells, slurries of paper pulp and coal, paint and even mud — are much messier than those made from monodisperse spheres, as the particles have a large distribution of shape, size and interparticle interactions. Nevertheless, some of the non-newtonian visco-elastic phenomena shown by natural colloids, such as shear thinning, can also be modelled by studying monodisperse colloids.

If the density of spheres is high, they interact strongly, and they can even crystallize if the size variation is less than about five per cent. This crystallization gives opals their fire, as visible light is diffracted from the ordered lattice planes of silica spheres³; and the same mechanism causes the crystal-

lization of proteins for structural determination. It is the effective particle-particle interaction potential that determines the phases and phase transitions of these condensed monodisperse suspensions, as well as those of the messier colloids mentioned above.

Unfortunately, for highly charged colloids suspended in an electrolyte, the effective pair potential or mean force between particles is uncertain⁴. The nature of the interactions is still disputed after nearly 60 years of experiment and theory because the experimental evidence has been indirect, mostly from polydisperse suspensions, and the theory is complicated by many-body effects. Moreover, even with monodisperse, spherical colloidal particles the suspensions are strongly inhomogeneous liquids, with up to 100,000 molecular ions screening the



Crystals that shouldn't last. A long-range attraction between like-charged particles explains the unexpected stability of these colloidal crystallites.

charge on each macro-ion. But Larsen and Grier now have experimental proof that the effective pair potential between two like-charged polystyrene spheres in water electrolyte includes a long-range attraction. This counter-intuitive result only occurs at high-to-intermediate sphere density, however, and with highly charged spheres, in electrolytes with large screening lengths.

How can two similarly charged particles attract each other? The established theory of colloidal particle interactions^{5,6} tells us that, at these separations, the screened Coulomb interactions should dominate the van der Waals dipole interactions. That means that the pair potential between Larsen and Grier's spheres should be everywhere repulsive. Indeed, when, using 'optical tweezers'⁷, they place two spheres a few diameters apart from each other (and well separated from all others and far from the glass plates), Larsen and Grier find a repulsive force. But when they put two spheres near a charged glass wall, they find a net attraction.

There is evidence for attraction within crystals too. To create the superheated crystals they need, the authors use an electrophoretic pulse to push all of the spheres near to one of the two parallel glass plates. Because of the roughly twofold density increase near this plate, several layers of face-centred cubic crystallites form there, with nearest-neighbour separation of about three diameters. With high salt content or small screening lengths, when the external electrical field is removed these crystallites fall apart within the time of a few video frames, less than a tenth of a second, as expected for purely repulsive interactions. But at low salt concentration and larger screening lengths the crystals are stabler — the sign of an attractive potential — and last for up to 10 minutes. They are also clearly faceted, which implies that the latent heat for freezing per particle is greater than $8kT$, already a factor of at least eight larger than expected for repulsive interactions.

What is the cause of this attractive force? Why is it only observed at high sphere density and long screening length where intuitively one would expect greater repulsion? The old theory neglects any effect that a charged sphere has on the screening ion density in the vicinity of other spheres — a nonlinear effect that will be largest at high particle density and great screening length. It seems that the attraction between two spheres seen in the new experiments is caused by a nearby charged double layer of screening ions next to a lattice plane of other spheres or a charged glass wall. But the details of the mechanism aren't clear.

The implications of the Larsen and Grier paper are that a simple pairwise additive interaction potential is not a good approximation for charged colloidal particles at high densities and large screening lengths. Differ-

ent indirect measurements of the interaction potential will depend on the measurement technique and approximations involved, which explains some of the confusion in the literature. The effective mean force between particles may even be dependent on the exact spatial distribution of particles, and thus, if there are nearby screening clouds due to other macro-ions, on direction.

Could this attraction explain the flocculation of some vats of paint or the formation of obstructing plugs in the flow of slurries, even when purified? Time will tell. Certainly,

attractive effects are industrially important in the stability of suspensions.

Cherry A. Murray is in the Semiconductor Physics Research Department, Bell Labs, Lucent Technologies, 1D-334, 700 Mountain Avenue, Murray Hill, New Jersey 07974-0636, USA.

1. Larson, A. E. & Grier, D. G. *Nature* **385**, 230–233 (1997).
2. Pieranski, P. *Contemp. Phys.* **24**, 25–73 (1983).
3. Sanders, J. V. *Phil. Mag. A* **42**, 705–720 (1980).
4. Chu, X. & Wasan, D. T. *J. Colloid Interface Sci.* **184**, 268–278 (1996).
5. Derjaguin, B. V. & Landau, L. *Acta Physicochimica* **14**, 633–662 (1941).
6. Verwey, E. J. & Overbeek, J. T. *Theory of the Stability of Lyophobic Colloids* (Elsevier, Amsterdam, 1948).
7. Crocker, J. & Grier, D. *Phys. Rev. Lett.* **73**, 352–355 (1994).

Glycolysis

Phosphotransfer hinges in PGK

Colin Blake

Biochemists of the future may wonder what the fuss with phosphoglycerate kinase (PGK) was all about. After all, although PGK is an essential enzyme in both photosynthesis and glycolysis, it is a relatively small monomeric molecule that catalyses a reversible phosphotransfer reaction (see figure) from 1,3-bisphosphoglycerate to ADP to form ATP and 3-phosphoglycerate (3-PGA). Yet it has resisted more than 20 years of kinetic and structural investigation. Now on page 275, Bernstein *et al.*¹ report that the crystal structure of a fully folded ternary complex of *Trypanosoma brucei* PGK has finally been solved — and this seems to pull together all of the previous observations and hypotheses.

The main difficulties in studying the behaviour of PGK arose because we needed to understand how its two substrates, bound to widely separated domains, could come into close contact for the direct phosphotransfer mechanism to operate (but only when both substrates are present). It was suspected that the two-domain structure of PGK was important; however, this was not confirmed until relatively recently, when the two domains were found to map almost exactly to the amino- and carboxy-terminal halves of the polypeptide chain². The binding of the Mg-ATP/ADP substrate was demonstrated relatively easily, using crystals grown from ammonium sulphate, but the second substrate (3-PGA) proved to be more of a problem.

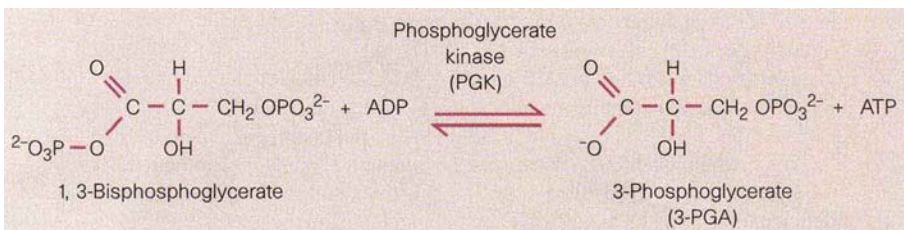
Assuming that the binding of 3-PGA was inhibited by the high concentrations of sulphate in the crystallization mixture, my group² identified a cluster of arginine and histidine residues — subsequently called the 'basic patch' — that offered a suitable binding site for the negatively charged phosphoglycerates. Although the site was located on the face of the amino-terminal domain, opposite the ATP-binding site on the carboxy-terminal domain, the site was too far away for the direct phosphotransfer of the type known to occur in PGK. We concluded that the only solution was a substantial 'hinge-bending'

conformational change to bring the two domains, and more particularly their bound substrates, together in the catalytic cycle.

At almost the same time, X-ray solution-scattering studies showed that the radius of gyration of ternary (Mg-ADP-PGK-3-PGA) complexes was markedly lower than for binary complexes or for the native enzyme³, consistent with the substrate-induced hinge-bending hypothesis. But little further progress was made until PGK crystals were grown from non-ionic media by Karl Harlos, Maria Vas and myself⁴. We were able to confirm that the basic patch was indeed the binding site for 3-PGA, and we also showed that 3-PGA causes a partial closing of the hinge.

Bernstein *et al.*¹ have now generated crystals of the *T. brucei* Mg-ADP-PGK-3-PGA complex. These crystals show a domain rotation of 32° with respect to the native horse enzyme — a transition that involves atom shifts, between the native enzyme and the complex, of up to 27 Å. Because a significant conformational change is not observed when the PGK-Mg-ADP complex forms, and formation of the PGK-3-PGA complex induces a domain rotation of only 8°, the new crystal can reasonably be seen to confirm the results of the solution-scattering study: that the large-scale hinge-bending change occurs only in the presence of both substrates.

How does PGK accomplish this? The picture that emerges from the ternary crystal structure is that the small conformational changes that are induced by the binding of 3-PGA 'prime' the 3-PGA-bound enzyme to react differently from the native enzyme to the binding of Mg-ATP/ADP. At the heart of the priming mechanism is helix 14, which is part of the amino-terminal domain. On binding 3-PGA, helix 14 undergoes a translational movement which, when Mg-ADP is bound, enables the hydrogen bonding in a partially formed β -sheet to be completed, thereby locking the amino- and carboxy-terminal domains together in a new orientation. The hinge-bending conformational change is sta-



Phosphoglycerate kinase (PGK) catalyses a reversible phosphotransfer reaction that is normally involved in the production of ATP during glycolysis. Bernstein *et al.*¹ have crystallized an active form of PGK, using substrates — ADP and 3-PGA — that cannot be converted to products. They find that when both substrates are bound, the enzyme undergoes a conformational rearrangement that brings the two ligand-binding sites together.

bilized by this mechanism, and allows the formation of a catalytically competent active site.

Hinge-bending requires the presence of both substrates, so the futile hydrolysis of 1,3-bisphosphoglycerate or ATP is prevented. But when both substrates are present, they are brought close enough for a direct phosphotransfer reaction to operate. Bernstein *et al.*¹ have shown that the key factors in the catalytic mechanism include the specific binding of each of the three oxygens of the planar phospho (PO_3) group of the transition state. The transferable phospho group is further stabilized by the helix dipole generated by helix 14.

The conformational change that is shown by this new crystal structure seems to answer the main questions about how PGK brings its two substrates into close proximity in a water-free catalytic site that does not exist in the native enzyme. Although the separate elements in this are not all new — the earlier

modelling of the conformational change accurately located the hinge and the magnitude of its rotation⁵, and the catalytic mechanism suggested here is similar to that proposed by May *et al.*⁶ — these experimental results integrate the subtle molecular changes, and provide the first detailed insight into the way in which PGK is fitted for its precise function in glycolysis.

Colin Blake was formerly in the Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU, UK.

1. Bernstein, B. E., Michels, P. A. M. & Hol, W. G. J. *Nature* **385**, 275–278 (1997).
2. Banks, R. D. *et al.* *Nature* **279**, 773–777 (1979).
3. Pickover, C. A., McKay, D. M., Engelman, D. M. & Steitz, T. A. *J. Biol. Chem.* **254**, 11323–11329 (1979).
4. Harlos, K., Vas, M. & Blake, C. C. F. *Proteins* **12**, 133–144 (1992).
5. Blake, C. C. F., Rice, D. W. & Cohen, F. E. *Int. J. Pept. Protein Res.* **27**, 443–448 (1986).
6. May, A., Vas, M., Harlos, K. & Blake, C. C. F. *Proteins* **24**, 292–303 (1996).

Stellar physics

Anaemia reveals star's great age

Mike Bolte

Astronomy is faced with a particularly intriguing dilemma over the age of the Universe. The current best estimate for the expansion time since the Big Bang is less than the measured age of the ancient globular clusters of our Galaxy. Now, for the first time, an independent age has been measured for a single halo star^{1,2}, and it deepens the difficulties that cosmology faces.

If the Universe is homogeneous and isotropic, its expansion age can be deduced from three observational quantities: the current expansion rate (the Hubble constant), determined by extragalactic distance measurements; the amount of matter in the Universe, which acts to slow down expansion; and the energy density of the vacuum (the cosmological constant), which exerts both accelerative and decelerative forces. Current measurements^{3–5} imply expansion ages of between 8.9 and 11.5 Gyr (billion years).

But the globular star clusters that live in the halo of our Galaxy appear to be older than that. Comparing the distribution of

luminosity and colour of the individual cluster stars with models of stellar structure and evolution has consistently led to cluster ages between 13 and 18 Gyr. There is little room for adjustment^{6–8}.

So who has something wrong — the cosmologists, the extragalactic distance-scale army, or the stellar modellers? Because they are in some ways the most complicated, the stellar models come in for most suspicion. Although they pass each observational test so far devised, before we can have full confidence in the age of the Universe as derived from stars, we must develop other techniques for measuring the ages of old stellar systems.

In the 1 May *Astrophysical Journal*, Cowan *et al.*¹ will present a simple, independent estimate of the age of one very old star, CS22892–052, using a particularly clean variation of a method called nucleocosmochronology. Usually in this method, the abundance of a long-lived radioactive element (or the ratio of two elements) is compared with its predicted initial value. That



100 YEARS AGO

The Monthly Weather Review

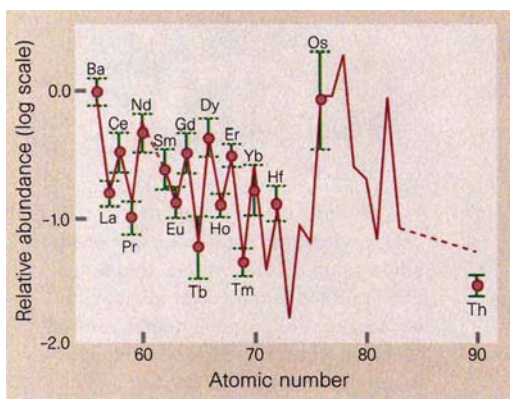
(Washington) for September last contains, among various other interesting notes, one upon the first attempt to measure wind force. Prof. Marvin points out that Sir Isaac Newton in his boyhood made a rough determination of the force of a great gale which occurred on September 3, 1658, by jumping first in the direction in which the wind blew, and then in opposition to the wind, and afterwards measuring the length of the leap in both directions. An account of this will be found in Sir David Brewster's "Memoir" of Sir Isaac Newton. The first piece of apparatus applied to the measurement of wind was probably the pendulous plate anemometer introduced by the Royal Society on the recommendation of Sir Christopher Wren and others, about 1665. This instrument gave a measurement of the effect of moving air on a resisting plate. The question of the measurement of the pressure or velocity of the wind by anemometers is still in a condition far from satisfactory, and the recent annual reports of the Meteorological Council show that the subject is still engaging the attention of that body.

From *Nature* 14 January 1897.

50 YEARS AGO

Statistical analyses have been made of the whole intake of the Science Library, London, not merely of three per cent samples. The results are for that reason more reliable, although there is still room for more research, for which financial aid would be desirable... It was established that about 3/4 million useful scientific and technical papers are recorded every year in an aggregate of some 15,000 useful scientific periodicals, while 300 abstracting and indexing periodicals publish also 3/4 million abstracts, or references. But these abstracts, or references, relate to only 1/4 million different papers. So that more than half the useful discoveries and inventions are recorded, only to be buried on the library shelves. This general analysis was confirmed by the examination of the literature of certain subjects. The details of the examination of electrical engineering literature show that only 11,500 different papers on this subject, out of an annual output of 24,000, are covered by a total of 45,000 abstracts, or references.

From *Nature* 18 January 1947.



Thorium loss. The pattern of heavy-element abundances in the star CS22892-052 (green) matches that in the Sun (yellow line), except in the case of ^{235}Th , which has a radioactive half-life of about 14 billion years. CS22892-052 is probably between 13 and 21 billion years old. (From ref. 1.)

prediction requires models for the chemical evolution of the Galaxy and for the complicated chains of element synthesis in stars and supernovae. Although there have been careful investigations along these lines before, the results have been considered slightly dubious because of the strong dependence on modelling complicated processes. To see how the new work bypasses these uncertainties requires a little background.

The bulk of every element heavier than boron has been formed by nuclear reactions in stars, becoming recycled into the interstellar medium and incorporated into new stars. That is why the oldest stars in the Galaxy are much poorer in these elements than is the Sun. Equilibrium nuclear fusion produces elements up to iron; the heavier elements are mostly formed by non-equilibrium reactions, where neutrons are added to nuclei to build heavy, unstable atoms which then decay into stable atoms. One such route is the 'r-process', in which many neutrons are added between decays.

Nearly all the potential 'nucleochronometers' — elements that have half-lives in the region of one to ten Gyr — are produced via the r-process. But there is an astronomical catch-22 involved in measuring the abundances of most r-process elements. Their generally weak absorption lines can be overwhelmed by a forest of other line blends — iron is the biggest culprit — and attempting to minimize the iron confusion by observing more chemically deficient stars doesn't necessarily work, because the lines from the elements of interest are likely to be weaker too.

Fortunately, there is a way out. The chemically poor stars of the Galactic halo can have iron abundances as small as 1/10,000 of the solar value, but the ratio of r-process elements to iron is usually higher than it is in the Sun. Why? The r-process is thought to occur in very massive stars, which explode as supernovae after a short life. In contrast, much of the iron in the Galaxy is thought to

be produced in supernovae of a different type, with much longer-lived progenitor stars; so the old halo stars were formed before there was much iron around.

This is where CS22892-052 enters the story. It is one of the 'ultra-metal-deficient' stars in the halo, with the highest-known ratio of r-process elements to iron — so high that it is possible to measure the abundances of 17 such elements, from barium to thorium. As first reported in Sneden *et al.*² and discussed more fully by Cowan *et al.*¹, when the abundances of stable r-process elements in CS22892-052 are plotted against scaled Solar System abundances, the patterns match almost perfectly (see figure). The only exception is ^{235}Th , which in CS22892-052 sits a factor of two

below the scaled solar value (corrected for decay over the past 4.6 Gyr).

Based on the pattern of stable elements, it seems inescapable that the r-process synthesis that preceded the formation of CS22892-052 was essentially identical to the one that preceded the formation of the Sun. Given the factor of two deficiency in thorium, and its half-life of 14.2 Gyr, it is trivial to derive an age of about 15 Gyr. This estimate is independent of Galactic chemical evolution and of stellar-model-based ages of globular clusters.

But this is actually a lower limit to the age of CS22892-052, as it assumes that the thorium in the Solar System was synthesized just before the formation of the Sun. In fact, much of it is bound to be older, so the initial thorium abundance was higher and the age of CS22892-052 must be greater. The authors' preferred evolution model corrects the age of this star to 17 ± 4 Gyr.

Does this prove the stellar modellers correct? With only this single measurement in one star, and the single line of thorium, it is premature to consider the age of the oldest stars a solved problem. Observations with improved signal-to-noise ratio and higher spectral resolution are possible, however, and will decrease the large uncertainty. There are many additional candidate stars, and in the next few years this will be a hotly pursued line of research. But this is already an ominous discovery for conventional cosmology. It is rare to have such a clean result with bearing on such an important problem.

Mike Bolte is at the UCO/Lick Observatory, University of California, Santa Cruz, California 95064, USA.

1. Cowan, J. J., McWilliam, A., Sneden, C. & Burris, D. L. *Astrophys. J.* (in the press).
2. Sneden, C. *et al. Astrophys. J.* **467**, 819–839 (1996).
3. Jacoby, G. *et al. Publ. Astron. Soc. Pacif.* **104**, 599–662 (1992).
4. van Den Bergh, S. *Science* **270**, 1942–1943 (1995).
5. Farreese, L. *et al. Astrophys. J.* **464**, 568–599 (1996).
6. Bolte, M. & Hogan, C. J. *Nature* **376**, 399–402 (1995).
7. Chaboyer, B. *Astrophys. J. Lett.* **444**, L9–L12 (1995).
8. Vanden Berg, D. A. *et al. Annu. Rev. Astron. Astrophys.* **34**, 461–510 (1996).

Daedalus

Youth in age

Spare-part surgery is a demanding business. Organs such as hearts, livers, kidneys, and so on, can often be implanted successfully into patients whose own organs have failed. But nervous and glandular tissue is very hard to replace, and rejection is always a problem. Daedalus now has a new approach.

Every part of our bodies is formed by the differentiation of fetal cells, guided by the environment around them. By the time a part is finally formed, its cells have more or less exhausted their regenerative capacity, and can seldom repair serious damage. But imagine, says Daedalus, that a small amount of (e.g.) brain tissue were removed from a newborn baby, and stored in liquid nitrogen. Many decades later, its owner might begin to suffer brain trouble: perhaps Alzheimer's or Parkinson's disease. The long-stored sample of tissue could be warmed up, and cultured in a suitable medium (even brain cells can be cultured these days). The result would be a useful quantity of the patient's own brain cells. They could be injected into the site of damage without fear of rejection. And these new cells would still have all the plasticity of youth. They would differentiate in the normal youthful way, taking their clues from the cells around them. They would make the right sort of connections to the rest of the brain, and form a splendid 'hardware patch' for the failing region.

This scheme can clearly be generalized. Small samples of many crucial tissues — liver, nerve, bone marrow, thyroid, gonad, and so on — could be taken from each newborn baby, and stored in liquid nitrogen as an insurance against future trouble.

Accident and old age would lose many of their terrors. No longer would their victims have to wait for some donor of a new part to become available. Their own repair material would be always at hand, ready to grow spontaneously into the required form on simple injection. Even the normal but still resented features of advancing age, such as sexual decline, might be alleviated. At the age of perhaps 35 or 40, one might have a stored sample of one's own neonatal gonadal tissue reinjected. Nothing much would happen; but ten or twelve years later, the injected tissue would start its own private pubertal development. Its owner would enjoy a surge of new youthful vigour and enterprise, countering the normal decline. He or she would probably not live any longer, but would enjoy a much livelier old age. And governments would be delighted. They would be able to reset pensionable age to about 95.

David Jones

Computation and the single neuron

Christof Koch

Neurons and their networks underlie our perceptions, actions and memories. The latest work on information processing and storage at the single-cell level reveals previously unimagined complexity and dynamism.

Over the past few decades, neural networks have provided the dominant framework for understanding how the brain implements the computations necessary for its survival. At the heart of these networks are simplified and static models of nerve cells. But neuroscience is undergoing a revolution, and one consequence is that the picture of how neurons go about their business has altered.

To appreciate that, we need to start with a simple account of neuronal information-processing. A typical neuron in the cerebral cortex, the proverbial grey matter, receives input from a few thousand fellow neurons and, in turn, passes on messages to a few thousand other neurons (Fig. 1). These connections are hard-wired in the sense that each connection is made by a dedicated wire, the axon, so that, unlike processors in a computer network, there is no competition for communication bandwidth.

Apart from axons, there are three other principal components of a neuron. The cell body (or soma); dendrites, which are short, branched extensions from the cell body, and which in the traditional view simply receive stimuli from other neurons and pass them on to the cell body without much further processing; and synapses, the specialized connections between two neurons.

Synapses are of two types, excitatory and inhibitory. An excitatory synapse will slightly depolarize the electrical potential across the membrane of its target cell, while an inhibitory input will hyperpolarize the cell. If the membrane potential at the cell body exceeds a certain threshold value, the neuron generates a millisecond-long pulse, called an action potential or spike (Fig. 2a, overleaf). Otherwise, it remains silent. The amount of synaptic input determines how fast the cell generates spikes: these spikes are in turn conveyed to the next target cells through the output axon. Information processing in an average human cortex would rely on the proper interconnection of about 4×10^{10} such neurons in a network of stupendous size.

In 1943, McCulloch and Pitts showed that this view is at least plausible. They made a series of simplifications that have been with us ever since. Mathematically, each synapse is modelled by a single scalar weight, ranging

from positive to negative depending on whether the synapse is excitatory or inhibitory. As a whole, the neuron is represented as a linear threshold element; that is, the contributions from all synapses, multiplied by their synaptic weights, add linearly at the cell body. If the threshold is exceeded, the neuron generates a spike. McCulloch and Pitts argued that, with the addition of memory, a sufficiently large number of these logical 'neurons', wired together in an appropriate manner, can compute anything that can be computed on any digital computer.

Significant as their result was, it left some major questions unanswered. Most importantly, how is such a network set up in the absence of an external programmer? A network of these abstract neurons has to be programmed from the outside to do any particular job, but the brain assembles by itself. Clearly, the information needed to provide a unique specification of the 2×10^{14} synapses in human cerebral cortex cannot possibly be encoded in the genome. Moreover, a key feature of biological nervous systems that radically distinguishes them from present-day computers is their ability to learn from previous experience, presumably by adjusting synaptic weights. Finally, the properties of real neurons are much more elaborate and variable (Fig. 2b) than the simple and reliable units at the heart of McCulloch and Pitts's neural networks and that of their progeny.

A possible mechanism addressing the

need for self-programming was postulated a few years later by Donald Hebb¹: "When an axon of cell A is near enough to excite cell B and repeatedly or consistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased". Here, according to the principle known as synaptic plasticity, the synapse between neuron A and neuron B increases its 'weight' if activity in A occurs at the same time as activity in B.

Much of the excitement in the neural-network revolution of the 1980s was triggered by the discovery of learning rules for determining the synaptic weights using variants of Hebb's rule for synaptic plasticity². These rules allow the synaptic weights to be adjusted so that the network computes some sensible function of its input — it learns to recognize a face, to encode eye position or to predict the stock market, for example.

Underlying these developments was the view that memory is stored in the synaptic weights. Changing connection strength can make the network converge to a different stable equilibrium, equivalent to recalling a different memory (associative memory; ref. 3). Experimentally, the best-studied aspect of increasing connection strength is long-term potentiation, which can be conceptualized as an increase in synaptic weight lasting for days or even weeks. It is induced by simultaneous activity at the pre- and postsynaptic terminals (Fig. 1), in agreement with Hebb's rule⁴. Of more recent vintage is the discovery of a complementary process, a decrease in synaptic weight termed long-term depression⁵.

These neural-network models assumed that, to cope with the apparent lack of reliability of single cells, the brain makes use of a 'firing-rate' code. Here, only the average number of spikes within some suitable time window, say a fraction of a second, matters. The detailed pattern of spikes was thought to be largely irrelevant (Fig. 2b). This hypothesis was supported by experiments in which

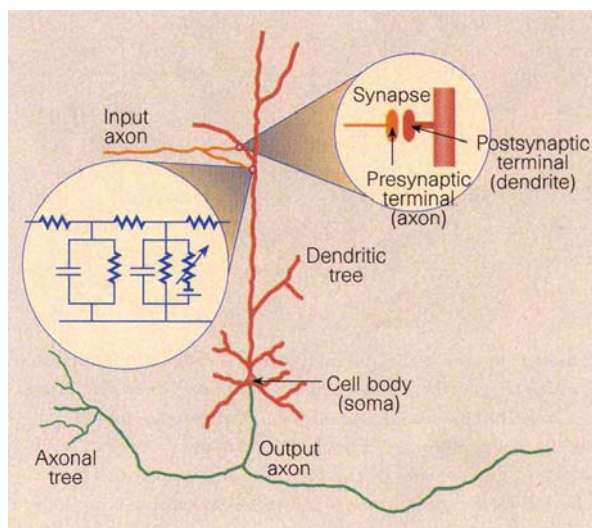


Figure 1 Drawing of a cortical pyramidal cell, a typical neuron. Incoming axons from other neurons connect with synapses on the dendritic tree. The electrical signals are then conveyed to the cell body, where action potentials (spikes) are generated. These are passed along the output axon to target neurons. The inset at left depicts a standard electrical model of a passive dendrite (modelled as a set of resistance-capacitance elements), with a synapse (modelled as a battery in series with a time-dependent conductance).

monkeys were given the task of discriminating a pattern of moving dots amid background noise⁶. The monkey's performance could be statistically predicted by counting spikes in single neurons within visual cortex, implying that all of the information needed to perform the task is available from the average firing rate alone. Further precision could in principle be obtained by averaging the firing rates over a large number of neurons (a process known as population coding), without any need to invoke the precise timing of individual spikes as a source of information.

A snapshot of the 'standard' view of information processing in the brain, say as of 1984, would be based on simple, linear-threshold neurons using a firing-rate code that waxes and wanes over hundreds of milliseconds to communicate information among neurons. Memory and computation are assumed to be expressed by the firing activity of large populations of neurons whose synaptic weights are appropriately set by synaptic learning algorithms. But since then neuroscience has undergone a phase of exponential growth, with progress in three separate areas — in the study of dendrites, of spikes and their timing, and of synaptic plasticity.

Active dendrites

It was Ramón y Cajal who, in the late nineteenth century, first revealed the intricacies and complexities of real neurons. He showed that a prototypical neuron, say a cortical pyramidal cell as shown in Figs 1 and 3, has an extended dendritic tree which is scores of times larger in surface area than the cell body. It has become clear that dendrites do much more than simply convey synaptic inputs to the cell body for linear summation. Indeed, if this were all they did, it is not obvious why dendrites would be needed at all; neurons could be spherical in shape and large enough to accommodate all the synaptic inputs directly onto their cell bodies. A few neurons do follow this geometrical arrangement but the vast majority are more like the cell shown

in Fig. 3, with an extended dendritic tree. This is where many of the synaptic inputs to the cell are received, but much of the membrane area is still devoid of synapses; so the function of this elaborate structure cannot simply be to maximize the surface area for synaptic contact.

Dendrites have traditionally been treated as passive cables, surrounded by a membrane which can be modelled by a conductance in parallel with a capacitance⁷. When synaptic input is applied, such an arrangement acts as a low-pass filter, removing the high frequencies but performing no other significant information processing. Dendrites with such a passive membrane would not really perturb our view of neurons as linear threshold units.

But dendrite membranes are not passive — they contain voltage-dependent membrane conductances. These conductances are mediated by protein complexes that allow charged ions to flow across the membrane. As long ago as the 1950s, Hodgkin and Huxley showed how the transient changes in such conductances generate and shape the action potential. But it was assumed that they are limited to the axon and the adjacent cell body. We now know that many dendrites of pyramidal cells are endowed with a relatively homogeneous distribution of sodium conductances as well as a diversity of calcium membrane conductances⁸.

What is the function of these active conductances? In a passive cable structure, synaptic input to the more distant regions of the dendritic tree (at the top in Fig. 3) would quickly saturate, delivering only a paltry amount of electric current to the spike-initiation zone far away. But it turns out that synaptic input to this part of the tree is sufficient to elicit somatic action potentials⁹, and the likely explanation (supported by computer models¹⁰) is that calcium and potassium membrane conductances in the distant dendrites can selectively amplify this input.

Voltage-dependent conductances can

also subserve a specific nonlinear operation, multiplication. In a passive dendritic tree, the effect of two synaptic inputs is usually less than the sum of the two individual inputs; that is, they show saturation. This saturation effect can be minimized by spreading the synapses far apart. If, however, the dendritic tree contains sodium or calcium conductances, or if the synapses use a particular type of receptor (the so-called NMDA receptor), the inputs can interact synergistically: now, the strongest response occurs if inputs from different neurons are located close to each other on a patch of dendritic membrane. Computer simulations¹¹ show that such a neuron effectively performs a multiplication; that is, its firing rate is proportional to the product, rather than the sum, of its inputs. Multiplication is one of the most common operations carried out in the nervous system (for example, for estimating motion or the time-to-contact with an approaching stimulus), and it is tempting to speculate that this may be achieved through such processes at the cellular level¹².

A further development has been the realization that the distribution of calcium ions within dendrites may represent another crucial variable for processing and storing information. Calcium enters the dendrites through voltage-gated channels, and this, along with its diffusion, buffering and release from intracellular stores, leads to rapid local modulations of calcium concentration within the dendritic tree. The concentration of calcium can, in turn, influence the membrane potential (through calcium-dependent membrane conductances) and — by binding to buffers and enzymes — turn local biochemical signalling pathways on or off.

It was also Ramón y Cajal who postulated the law of 'dynamic polarization', which stipulates that dendrites and cell bodies are the receptive areas for the synaptic input, and that the resulting output pulses are transmitted unidirectionally along the axon to its targets. This assumes that action potentials travel only along axons: no signal was thought to travel outwards along the dendrites.

From work on brain slices, however, it seems that this is by no means the whole story. Single action potentials can propagate not only forwards from their initiation site along the axon, but also backwards into the dendritic tree (a phenomenon known as 'antidromic spike invasion'¹³). It remains unclear whether dendrites can initiate action potentials themselves. If they can, such dendritic spikes could support theoretical proposals¹⁴ that all-or-none logical operations occur in the dendritic tree. The next step will be to find out whether action potentials can propagate into the dendritic tree under more natural conditions — that is, using sensory stimuli in an intact animal¹⁵.

Thus, it is now evident that the dendritic

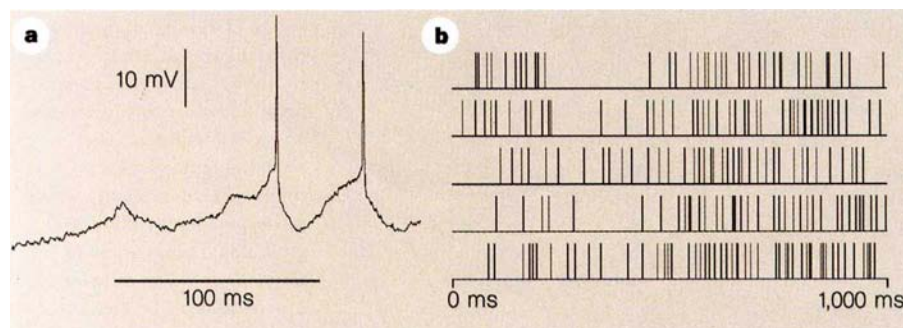


Figure 2 a, The membrane potential of a single neuron in the cat visual cortex. A visual stimulus causes a sustained barrage of synaptic input, which triggers three cycles of depolarization. The first cycle does not reach the threshold for generating a spike, but the second and third do. (Data provided by B. Ahmed and K. Martin.) b, Variability of the firing response of a single neuron in monkey cortex (each line in the trace corresponds to a spike of the type shown in a). The same stimulus is presented five times and triggers about 40 spikes each time; yet the exact timing of individual spikes shows random variation. (Data provided by W. Newsome and K. Britten.)

tree is far more complex than the linear cable models of yesteryear assumed. Dendrites provide the substrate for numerous non-linear operations, and endow neurons with much greater information-processing capacity than was previously suspected.

Timing counts

The second area in which our thinking has changed has to do with the role of time in neuronal processing. There are two main aspects to this issue — first, the relationship between the timing of an event in the external world and the timing of the representation of that event at the single-neuron level; second, the accuracy and importance of the relative timing of spikes between two or more neurons.

Regarding the first question, some animals can discriminate intervals of the order of a microsecond (for instance to localize sounds), implying that the timing of sensory stimuli must be represented with similar precision in the brain. But this usually involves highly specialized pathways, probably based on the average timing of spikes in a population of cells. However, it is also possible to measure the precision with which individual cells track the timing of external events. For instance, certain cells in the monkey visual cortex are preferentially stimulated by moving stimuli, and these cells can modulate their firing rate with a precision of less than 10 ms (ref. 16).

The second aspect of the timing issue is the extent to which the exact temporal arrangements of spikes — both within a single neuron and across several neurons — matters for information processing. In the past few years there has been a resurgence in signal-processing and information-theoretical approaches to the nervous system¹⁷. In consequence we now know that individual neurons, such as motion-selective cells in the fly or single auditory inputs in the bullfrog, can encode between 1 and 3 bits of sensory information per output spike, amounting to rates of up to 300 bits per second. This information seems to be encoded using changes in the instantaneous interspike interval between a handful of spikes. Such a temporal encoding mechanism is within 10–40 per cent of the theoretical maximum allowed by the spike-train variability. This is quite remarkable because it implies that individual spikes in a single cell in the periphery can carry significant amounts of information, quite at odds with the idea that neurons are very unreliable and can only signal in the aggregate. At these rates, the optic nerve, which contains about one million fibres, would convey between one and a hundred million bits per second — compare this with a quad-speed CD-ROM drive, which transfers information at 4.8 million bits per second.

The relative timing of spikes in two or more neurons can also be remarkably pre-

cise. For instance, spikes in neighbouring cells in the lateral geniculate nucleus, which lies midway between the retina and visual cortex, are synchronized to within one millisecond (ref. 18). And within the cortex, pairs of cells may fire action potentials at predictable intervals that can be as long as 200 ms, two orders of magnitude longer than the delay due to a direct synaptic connection, with a precision of about 1 ms (ref. 19). Such timing precision across populations of simultaneously firing neurons is believed to be a key element in neuronal strategies for encoding perceptual information in the sensory pathways^{20–22}.

If neurons care so much about the precise timing of spikes — that is, if information is indeed embodied in a temporal code — how, if at all, is it decoded by the target neurons? Do neurons act as coincidence detectors, able to detect the arrival time of incoming spikes at the millisecond or better resolution? Or do they integrate more than a hundred or so relatively small inputs over many tens of milliseconds until the threshold for spike initiation is reached (ref. 23; Fig. 2a)? These questions continue to be widely and hotly debated¹⁴.

Synaptic plasticity

Back-propagating action potentials are puzzling if considered solely within the context of information processing. But they make a lot of sense if seen as 'acknowledgement signals' for synaptic plasticity and learning. A Hebbian synapse is strengthened when pre- and postsynaptic activity coincide. This can occur if the presynaptic spike coincides with the postsynaptic spike that is generated close to the cell body and spreads back along the dendritic tree to the synapse. A new and beautiful study²⁴ shows that the order of the arrival time between the presynaptic spike and the back-propagated postsynaptic spike is critical for synaptic plasticity.

If the presynaptic spike precedes the postsynaptic spike, as should occur if the first participates in triggering the second, then long-term potentiation occurs — that is, the synaptic weight increases. If, however, the order is reversed, the synaptic weight decreases. So sensitive is this sequence-dependence that a change in timing of as little as 10 ms either way can determine whether a synapse is potentiated or depressed. The purpose of this precision is presumably to enable the system to assign credit to those synapses that were actually responsible for generating the postsynaptic spike.

These experiments come at a fortuitous time, for theoretical work has begun to incorporate asymmetric timing rules into neural-network models. The attraction of doing so is that it allows the network to form associations over time, enabling it to learn sequences and to predict events²⁵. It is therefore gratifying to find that synapses with the required

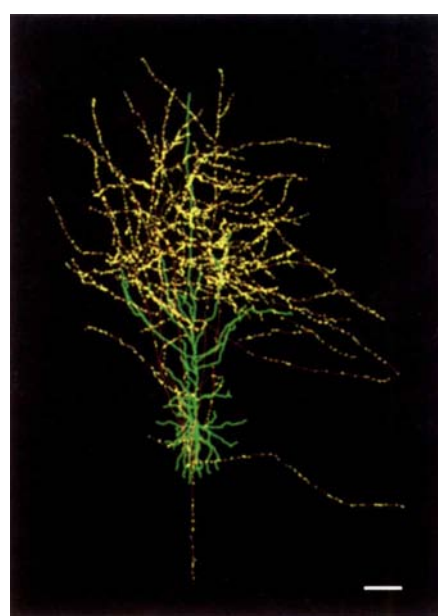


Figure 3 A pyramidal cell from cat visual cortex. It was injected with a tracer dye and reconstructed from serial light-microscope sections, revealing an unprecedented level of detail (and beauty). Its axon, projecting out of visual cortex, reveals a dense arborization (shown in red) and forms 4,105 synapses onto postsynaptic targets (yellow). The dendrites of the cell are shown in green. The scale bar corresponds to 100 μm . Some 100,000 cells are packed into 1 mm^3 of cortical tissue, which also includes about 4 km of axonal wiring, 500 m of dendrites and close to one billion synapses³⁰. (Data provided by J. Anderson, K. Martin and T. Binzegger; a stereoscopic image of this cell can be viewed at http://www.ini.unizh.ch/stereo_pict/stereo.html).

properties do indeed exist within the brain.

In the past year there has also been the emergence of a new way of thinking about short-term synaptic plasticity, one that complements the view of long-term synaptic changes for memory storage. This has come about by joint experimental-theoretical work (refs 26, 27; see box overleaf), suggesting that individual synapses rapidly adapt to the presynaptic firing rate, primarily signalling an increase or a decrease in their input. That is, synapses continuously adapt to their input, only signalling relative changes, which means that the system can respond in a highly sensitive manner to a constantly and widely varying external and internal environment. This is entirely different from digital computers that enforce a strict segregation between memory (on-board cache, RAM or disk) and computation. Indeed, they are carefully designed to avoid adaptation and other usage-dependent effects from occurring.

Interestingly, single-transistor learning synapses — based on the floating-gate concept underlying erasable programmable ROM digital memory — have now been built in a standard CMOS manufacturing

The adaptable synapse

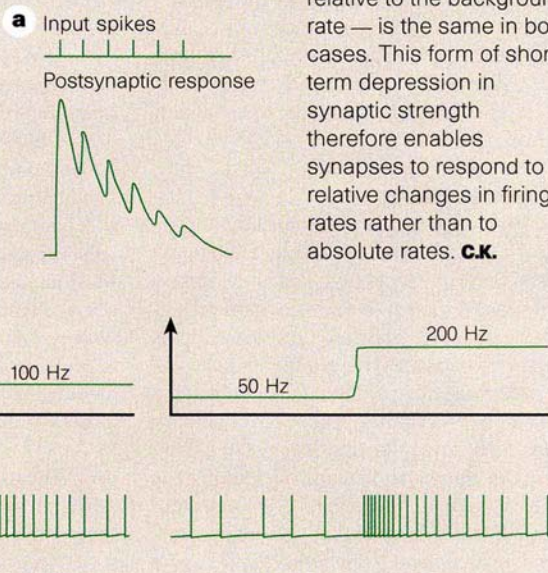
In short-term synaptic depression, the postsynaptic response to a regular train of presynaptic spikes firing at a fixed frequency f gradually lessens. The response to the first spike might be large, but subsequent responses will be diminished until they reach a steady state (expressed in terms of A , the fractional reduction in postsynaptic effect). This is shown here (part a) for presynaptic spikes at 40 Hz frequency (data from ref. 26).

This depression generally recovers within 0.1 to 0.5 s. For firing rates above 10 Hz, A is roughly inversely proportional to the firing frequency. In other words, within a few hundred milliseconds the synapse will have adapted to the

presynaptic firing with a response roughly independent of the firing rate (due to the inverse relationship between A and f). If the synaptic input now abruptly increases by δf , the adapted cell will initially increase its response by $\delta f/A$ which is proportional to $\delta f/f$. As a consequence, the transient change in the postsynaptic response will be proportional to the relative change in firing frequency. This is demonstrated in

computer simulations in which the presynaptic firing rate to a couple of hundred such synapses converging onto a model neuron is increased fourfold (part b — the increases are from 25 to 100 Hz on the left and from 50 to 200 Hz on the right; data from ref. 27).

Even though the final input rate is twice as high on the right side as on the left, the firing rate of the neuron is roughly the same. This is because the fractional increase — relative to the background rate — is the same in both cases. This form of short-term depression in synaptic strength therefore enables synapses to respond to relative changes in firing rates rather than to absolute rates. **C.K.**



process²⁸. Like biological synapses, they can change their effective weight in a continuous manner while they carry out computations. Floating-gate synapses will greatly aid attempts to replicate the functionality of nervous systems by the appropriate design of neuromorphic silicon neurons using analog very-large-scale-integrated (VLSI) circuit-fabrication technology²⁹.

Hybrid computer

Overall, then, current thinking about computation in the nervous system has the brain as a hybrid computer. Individual nerve cells convert the incoming streams of digital pulses into spatially distributed variables, the postsynaptic membrane potential and calcium redistribution. This transformation involves highly dynamic synapses that adapt to their input.

Information is then processed in the analog domain, using a number of linear and nonlinear operations (multiplication,

saturation, amplification, thresholding) implemented in the dendritic cable structure and augmented by voltage-dependent membrane and synaptic conductances. The resulting signal is then converted back into digital pulses and conveyed to the following neurons. The functional resolution of these pulses is in the millisecond range, with temporal synchrony across neurons likely to contribute to coding. Reliability could be achieved by pooling the responses of a small number (20–200) of neurons.

And what of memory? It is everywhere (but can't be randomly accessed). It resides in the concentration of free calcium in dendrites and the cell body; in the presynaptic terminal; in the density and exact voltage-dependency of the various ionic conductances; and in the density and configuration of specific proteins in the postsynaptic terminals.

Only very little of this complexity is reflected in today's neural-network literature. Indeed, we sorely require theoretical tools that

deal with signal and information processing in cascades of such hybrid, analog–digital computational elements. We also need an experimental basis, coupled with novel unsupervised learning algorithms, to understand how the conductances of a neuron's cell body and dendritic membrane develop in time. Can some optimization principle be found to explain their spatial distribution?

As always, we are left with a feeling of awe for the amazing complexity found in nature. Loops within loops across many temporal and spatial scales. And one has the distinct feeling that we have not yet revealed every layer of the onion. Computation can also be implemented biochemically — raising the fascinating possibility that the elaborate regulatory network of proteins, second messengers and other signalling molecules in the neuron carry out specific computations not only at the cellular but also at the molecular level.

Christof Koch is in the *Computation and Neural Systems Program*, California Institute of Technology, Pasadena, California 91125, USA (e-mail: koch@klab.caltech.edu).

1. Hebb, D. O. *The Organization of Behavior: A Neuropsychological Theory* (Wiley, New York, 1949).
2. Arbib, M. (ed.) *The Handbook of Brain Theory and Neural Networks* (MIT Press, Cambridge, MA, 1995).
3. Hertz, J., Krogh, A. & Palmer, R. G. *Introduction to the Theory of Neural Computation* (Addison-Wesley, Redwood City, CA, 1991).
4. Nicoll, R. A. & Malenka, R. C. *Nature* **377**, 115–118 (1995).
5. Stevens, C. F. *Nature* **381**, 471–472 (1996).
6. Newsome, W. T., Britten, K. H. & Movshon, J. A. *Nature* **341**, 52–54 (1989).
7. Segev, I., Rinzel, J. & Shepherd, G. *The Theoretical Foundation of Dendritic Function: Selected Papers of Wilfrid Rall with Commentaries* (MIT Press, Cambridge, MA, 1995).
8. Johnston, D., Magee, J., Colbert, C. & Christie, B. *Annu. Rev. Neurosci.* **19**, 165–186 (1996).
9. Caulier, L. J. & Connors, B. W. *J. Neurosci.* **2**, 751–762 (1994).
10. Bernander, O., Douglas, R. J. & Koch, C. *J. Neurophysiol.* **72**, 2743–2753 (1994).
11. Mel, B. W. *Neural Computation* **6**, 1031–1085 (1994).
12. Koch, C. & Poggio, T. in *Single Neuron Computation* (eds McKenna, T., Davis, J. & Zornetzer, S. F.) 315–345 (Academic, Boston, 1992).
13. Stuart, G. J. & Sakmann, B. *Nature* **367**, 69–72 (1994).
14. Softky, W. R. *Curr. Opin. Neurobiol.* **5**, 239–247 (1995).
15. Svoboda, K., Denk, W., Kleinfeld, D. & Tank, D. W. *Nature* **385**, 161–165 (1997).
16. Bair, W. & Koch, C. *Neural Computation* **8**, 1185–1202 (1996).
17. Rieke, F., Warland, D., van Steveninck, R. R. D. & Bialek, W. *Spikes: Exploring the Neural Code* (MIT Press, Cambridge, MA, 1996).
18. Alonso, J. M., Usrey, W. M. & Reid, R. C. *Nature* **383**, 815–819 (1996).
19. Abeles, M. *Corticonics: Neural Circuits of the Cerebral Cortex* (Cambridge Univ. Press, 1990).
20. Hopfield, J. J. *Nature* **376**, 33–36 (1995).
21. Singer, W. & Gray, C. M. *Annu. Rev. Neurosci.* **18**, 555–586 (1995).
22. DeCharms, R. C. & Merzenich, M. M. *Nature* **381**, 610–613 (1996).
23. Softky, W. R. & Koch, C. *J. Neurosci.* **13**, 334–350 (1993).
24. Markram, H., Lübke, J., Frotscher, M. & Sakmann, B. *Science* **275**, 213–215 (1997).
25. Montague, P. R., Dayan, P., Person, C. & Sejnowski, T. J. *Nature* **377**, 725–728 (1995).
26. Markram, H. & Tsodyks, M. *Nature* **382**, 807–810 (1996).
27. Abbott, L. F., Varela, J. A., Sen, K. & Nelson, S. B. *Science* **275**, 220–224 (1997).
28. Diorio, C., Hasler, P., Minch, B. A. & Mead, C. *IEEE Trans. Electron. Devices* **43**, 1972–1980 (1996).
29. Douglas, R. J., Mahowald, M. & Mead, C. *Annu. Rev. Neurosci.* **18**, 255–281 (1995).
30. Braitenberg, V. & Schüz, A. *Anatomy of the Cortex* (Springer, Berlin, 1991).

Pair bonding in spotted poison frogs

Once thought to interact very little with conspecifics, amphibians are now known to have elaborate mating systems¹. Pair bonding, defined as a social and copulatory relationship between two individuals that share some aspect of offspring rearing (K. Johnson and N. Burley, manuscript in preparation), has not previously been reported to occur in amphibians¹, although it has been suggested from observations of captive animals². I report here that the spotted poison frog, *Dendrobates vanzolinii*, an Amazon rainforest species, forms pair bonds and provides extended biparental care of the young under natural conditions, similar to some birds and mammals.

The poison-frog family Dendrobatidae, which is found in rainforests or cloud forests from southern Nicaragua to southern Brazil and Bolivia³, comprises about 165 species, classified in six genera. Nearly all species in the family have parental care of terrestrial eggs and tadpoles³. Parental care is more advanced in *Dendrobates*, in which several species feed their tadpoles unfertilized (nutritive) eggs^{4,5}.

During March and April 1996, I studied the social behaviour of *D. vanzolinii* in lowland tropical forest in the state of Acre, Brazil (latitude, 8° 16' S; longitude, 72° 46' W). Individual frogs can easily be recognized because each one has a unique pattern of bright yellow spots and/or short lines on a black background. I observed 12 pairs of frogs for a total of 50 hours and 44 min.

Figure 1 shows a pair of frogs of the species *D. vanzolinii* by a cavity in which their tadpole is developing. In addition to rearing a tadpole in the hole, the frogs may subsequently deposit and fertilize an egg which they attach to the inner wall of the cavity above the waterline. The newly developed tadpole is not allowed to slide into the water below because the tadpoles are cannibalistic. Instead, the male transports the new tadpole to another hole suitable for its development. The female does not accompany the male to the new cavity; after a few days he will guide her there by vocal communication.

Male frogs transported single tadpoles to tiny cavities (mean cavity size 3.0 × 1.7 cm and 17.9 cm deep) primarily in saplings and woody vines, at a mean height of roughly 1.2 m above the forest floor. Pairs cooperated extensively to feed their tadpoles. Males communicated with females by frequently repeated vocalizations. The male's call rate increased in the presence of a female, from 3.7 ± 0.3 (*n* = 5) to 7.5 ± 0.3 (*n* = 6) calls per min (mean ± s.e.m.; Mann-Whitney *U* test, *P* ≤ 0.001). If receptive, the female approached the male and began to follow him, sometimes as close



Figure 1 A pair-bonded male and female *D. vanzolinii*. The frogs have just emerged from a small cavity in a 5.5-cm-diameter sapling where they have undergone courtship behaviour, resulting in the deposition of nutritive (unfertilized) eggs for consumption by their tadpole developing in the hole. The female (left) is sitting at the entrance of the hole; the male has fully emerged and is carrying the pair's newly developed tadpole on his back. In general, dendrobatid frogs are small; male *D. vanzolinii* have a mean (± s.d.) snout-vent length of 17.0 ± 0.6 mm (*n* = 18), and females are 18.6 ± 0.7 mm (*n* = 12).

behind as 5–10 cm. The male guided the female through vines and small understorey branches to the cavity in which he had previously transported their tadpole. The male and female entered the cavity together, and the male's calls became softer. The pair remained for an extended period (roughly 3 h) inside the hole, during which time they underwent courtship behaviour, resulting in the deposition of unfertilized eggs for the tadpole to consume. Of 24 eggs collected, 14 were intact, and only two of these were fertilized. Normally, dendrobatids deposit terrestrial eggs; therefore, eggs deposited in water do not live, even if fertilized.

I never observed females entering holes unless accompanied by a male. Frogs did not feed their young every day, the average interval between feeding being almost 5 days, and I repeatedly saw the same males and females together throughout the study. I saw a male with two females in its territory only once. This cooperation between the same males and females in the same territory to transport and feed their tadpoles is an indication of pair bonding. The length of time a pair remains together is unknown,

but there was no evidence that pairs were less strongly bonded at the end of the study compared with the beginning.

Comparative studies are needed to elucidate the evolution of complex social behaviour of *Dendrobates* that feed their young nutritive eggs. In another lineage of *Dendrobates*, females, rather than males, transport tadpoles to small leaf axils and return periodically to deposit unfertilized eggs for their tadpoles unaccompanied by a male⁴. Deception of the female in the lineage with biparental care may have led to the evolution of uniparental care, but currently a well-supported phylogeny of the group does not exist to test this hypothesis.

Janalee P. Caldwell

Oklahoma Museum of Natural History and
Department of Zoology,
University of Oklahoma,
Norman, Oklahoma 73019, USA
e-mail: caldwell@ou.edu

1. Sullivan, B. K., Ryan, M. J. & Verrell, P. A. in *Amphibian Biology* Vol. 2 (eds Heatwole, H. & Sullivan, B. K.) Ch. 2 (Surrey Beatty, New South Wales, 1995).
2. Weygoldt, P. Z. *Zool. Syst. Evolut. Forsch.* **25**, 51–67 (1987).
3. Duellman, W. E. & Trueb, L. *Biology of Amphibians* (McGraw-Hill, New York, 1986).
4. Brust, D. G. *J. Herp.* **27**, 96–98 (1993).
5. Caldwell, J. P. in *Neotropical Biodiversity and Conservation* (ed. Gibson, A. C.) Ch. 4 (Mildred E. Mathias Botanical Garden, Los Angeles, 1996).

Carbon beads with protruding cones

While experimenting with the formation of vapour-grown carbon fibres, we have produced an apparently novel carbon structure consisting of a unique combination of rough-surfaced beads with protruding smooth cones (see Fig. 1). Transmission electron microscopy reveals that the beads are composed of crumpled sheets of graphene. The cones, in contrast, are composed of relatively straight sheets of graphene aligned parallel to the cone axis: Fig. 1b shows a low-resolution image of the tip of one of the cones. Visible at its centre is the hollow core of the catalytically grown filament, about 6 nm in diameter, on which the carbon has been deposited. The cone surface cuts across the graphene planes at an angle, leaving exposed plane edges (Fig. 1c).

Despite their dramatically different appearance, several experimental techniques, including X-ray diffraction, micro-Raman spectroscopy and electron energy-loss spectroscopy, show that the carbon making up the beads and the cones is

similar at the microscopic level, consistent with turbostratic graphene. The electron energy-loss analysis of the carbon K-edge showed a small difference in the plasmon energy, indicating that the smooth cones have a slightly higher electrical conductivity, probably because of their better structural order.

These structures were produced by chemical vapour deposition of carbon onto thin carbon filaments (~ 5 nm in diameter) made by a catalytic method^{1,2}. We carried out the deposition process under a flowing atmosphere containing 33% methane and 67% hydrogen for a duration of 160 min. (Argon gas was flowed through the reactor at all times when the temperature was elevated and these reactants were not present.) When this process is carried out at 1,000–1,150 °C, the common result is deposition of a relatively even layer of carbon that builds the diameter of the filament at a rate of about $10 \mu\text{m h}^{-1}$. When we raised the deposition temperature to 1,300 °C we then observed the formation of the bead and cone structure. Before, between and after the growth trials that produced the new structures we executed control experiments at 1,100 °C to verify that the reactor produced 'normal' deposition under this condition.

The mechanism by which the carbon has organized into two such different textures is not yet understood. One possibility is that two distinctly different deposition processes occur simultaneously, one of which uses the core filament as a template for relatively well-ordered carbon layers, and another which perhaps proceeds slightly faster and is less ordered, creating the crumpled sheet texture in beads. In this case, to explain the cone shapes, it would be necessary for each new layer of the well-ordered growth to start near the beads and progress outwards. This would result in a natural lag in the progression of growth from one layer to the next, creating the cone geometry.

Although it cannot be demonstrated conclusively from our images, it is plausible that such growth is occurring in the pattern of a helically wound scroll, a mode for which there is recent precedent³. Another possibility is that our deposition process has somehow produced a composite structure of well-ordered graphite surrounded by rough graphite, and that this structure is then etched by contaminant species (for example, oxygen) inside the oven as it cools down, to create the observed structure.

These novel carbon structures may have interesting applications. There has been substantial recent interest in filled nanotubes^{4,6}. The cones may be a convenient system for such studies to be carried out inside individual nanotubes. The material composing the body of the cone structurally

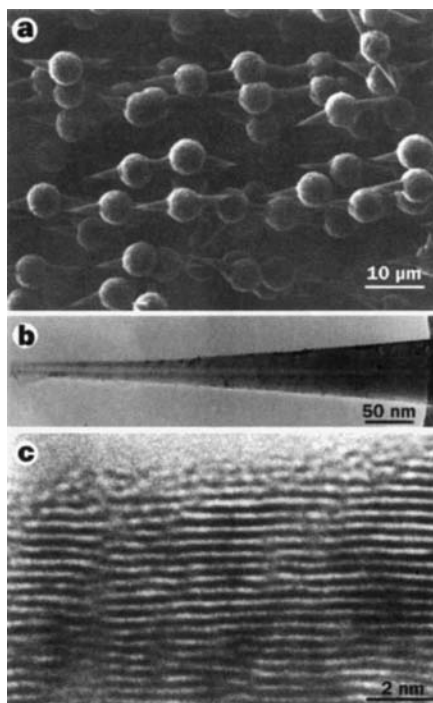


Figure 1 a, Scanning electron microscope image of carbon beads with protruding cones. b, Low-magnification transmission electron microscope image of a small cone. Note the hollow nanotube at the cone's core. c, Transmission electron microscope image near the surface of a cone. Note that the graphene planes are parallel to the cone axis (which is parallel to the bottom edge of the picture frame), and not the cone surface, leaving open plane edges.

reinforces the core nanotube making it mechanically robust, while the nearby beads could provide a grasping point for manipulation. Also, the shape of the cone makes the entry point of the nanotube easily located even under a normal scanning electron microscope. Another feature for exploitation may be the unique texture of exposed graphene edges at the cone surface. Such a characteristic could be useful for various purposes, such as grafting of chemical groups, bonding to a matrix, or use as a catalyst support.

Ronald L. Jacobsen

*Applied Sciences, Inc.,
Box 579, Cedarville,
Ohio 45314-0579, USA*

Marc Monthieux

*Centre d'Elaboration des Matériaux
et d'Etudes Structurales,
UPR A8011 CNRS,
BP 4347,
F31055 Toulouse Cedex, France*

1. Baker, R. T. K. *Carbon* **27**, 315–323 (1989).
2. Dresselhaus, M. S., Dresselhaus, G., Sugihara, K., Spain, I. L. & Goldberg, H. A. *Graphite Fibers and Filaments* 18–30 (Springer, Berlin, 1988).
3. Luyten, W. et al. *Ultramicroscopy* **49**, 123–131 (1993).
4. Dresselhaus, M. S. *Nature* **358**, 195–196 (1992).
5. Ajayan, P. M. & Iijima, S. *Nature* **361**, 333–334 (1993).
6. Tsang, S. C., Chen, Y. K., Harris, P. J. F. & Green M. L. H. *Nature* **372**, 159–162 (1994).

DNA analysis reveals the sex of infanticide victims

For many centuries, infanticide was an accepted practice for disposing of unwanted babies. We have obtained archaeological evidence of infanticide in Roman Ashkelon, on the southern coast of Israel, where skeletal remains of around 100 babies were discovered in a sewer^{1,2}. In ancient Roman society, infanticide was widespread and practised mostly against unwanted female babies³. We were therefore surprised when, by analysing ancient DNA, we found a significant number of male victims in Ashkelon.

We found the infants' remains beneath a bathhouse, built in the fourth century over earlier Roman villas, where lamps decorated with erotic images had previously been found¹. The Greek inscription "Enter, enjoy and..." which we discovered in the bathhouse, indicated that it might have also served as a brothel, as was common in the Roman empire⁴. The infants' bones were mixed with animal bones, potsherds and coins, in the gutter. Bone size, dental development and lack of neonatal lines in the teeth indicate that the human remains were 1- or 2-day-old infants. The combination of the death of so many babies showing no sign of disease or skeletal malformation, and the mode of disposal, implies infanticide².

There is ample evidence for infanticide of females in Graeco-Roman society^{3,5}. The most explicit reference comes from a letter in 1 BC, dated 17 June, written by a certain Hilarion to his wife Alis: "I ask and beg you to take good care of our baby son... If you are delivered of child... if it is a boy keep it, if a girl discard it"⁶.

It seems likely that many female babies were disposed of promptly after birth. We therefore thought that the Ashkelon bones would probably have been from females. The reliability of morphometric analyses for gender identification in infants is low, especially in the case of incomplete skeletons. Therefore we planned to deal with the gender question by DNA analysis of the X and Y alleles of the amelogenin gene.

Because studies of ancient DNA are prone to numerous artefacts⁷, stringent precautions and appropriate controls were required⁸. We extracted DNA from left femurs only, to avoid testing the same individual twice. We tested a total of 43 left femurs, and examined the reproducibility of the results in three independent extractions, using two different methods^{9,10}, and done several months apart. We performed three polymerase chain reactions (PCRs) for each DNA extract, using conditions and primers previously described¹¹.

DNA amplification was successful for 19

of the 43 specimens: 14 were found to be males and 5 females. The results for 17 of the specimens were reproduced in at least two separate DNA extracts. Significantly, there were no conflicting results in any of the different PCR analyses for a single specimen. We verified the authenticity of the amplified fragments by sequence analysis of male and female samples.

The significant number of male victims was unexpected and raised the intriguing possibility that these infants may have been the unwanted offspring of courtesans working in the bathhouse. This study exemplifies the usefulness of DNA analyses of human skeletal remains in obtaining unambiguous evidence to clarify otherwise open archaeological and anthropological questions.

Marina Faerman, Gila Kahila

Patricia Smith

Dental Division of Anatomy, Hebrew University,
Hadassah Faculty of Dental Medicine,
Jerusalem 91010, Israel

Charles Greenblatt

Department of Parasitology, Hebrew University,
Hadassah Medical School, Jerusalem 91010, Israel

Lawrence Stager

The Semitic Museum, Harvard University,
Cambridge, Massachusetts 02138, USA

Dvora Filon, Ariella Oppenheim

Department of Hematology, Hebrew University,
Hadassah Medical School, Jerusalem 91120, Israel

1. Stager, L. E. *Biblical Archaeol. Rev.* **17**, 38–57 (1991).
2. Smith, P. & Kahila, G. *J. Archaeol. Sci.* **19**, 667–675 (1992).
3. Harris, W. V. *J. Roman Studies* **84**, 1–22 (1994).
4. Ward, R. B. *Harvard Theol. Rev.* **85**, 125–147 (1992).
5. Eyben, E. *Ancient Soc.* **11/12**, 5–82 (1980–81).
6. Lewis, N. *Life in Egypt Under Roman Rule* 1–244 (Clarendon, Oxford, 1985).
7. Pääbo, S., Gifford, J. A. & Wilson, A. C. *Nucleic Acids Res.* **16**, 9775–9787 (1988).
8. Filon, D., Faerman, M., Smith, P. & Oppenheim, A. *Nature Genet.* **9**, 365–368 (1995).
9. Woodward, S. R., King, M. J., Chiu, N. M., Kuchar, M. J. & Griggs, C. W. *PCR Meth. Appl.* **3**, 244–247 (1994).
10. Hoss, M. & Pääbo, S. *Nucleic Acids Res.* **21**, 3913–3914 (1993).
11. Faerman, M. *et al. Gene* **167**, 327–332 (1995).

Medicinal alkaloid as a sex pheromone

The sex pheromones of more than 1,300 species of insect¹ have been identified since the milestone finding of bombykol more than 30 years ago, thanks to increasingly sophisticated analytical techniques and a great deal of effort. These sex pheromones, which number in their hundreds, are mainly restricted to a small group of chemicals with remarkable structural similarities. Female Lepidoptera, for example, largely use alcohols, aldehydes, acetates and hydrocarbons (including epoxides of hydrocarbons).

We have now identified an aromatic alkaloid, 1,3-dimethyl-2,4-(1*H*,3*H*)-quinazolinedione, as the sex pheromone for the

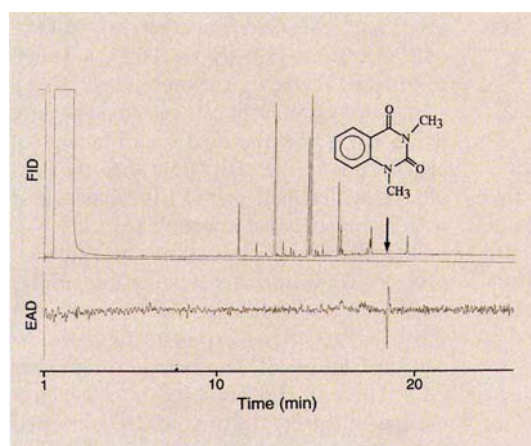


Figure 1 Parallel flame ionization detector (FID) and electroantennographic detector (EAD) chromatograms obtained from the injection of 5 female equivalents of the whole-body extract (ether) of female *P. diversa*. Chromatographic resolution was obtained on a BP-20 SGE capillary column operated at 100 °C for 1 min, rising 10 °C min⁻¹ to 260 °C, and held at this temperature for 10 min. A male antenna exposed to the effluent from a gas chromatograph responded only to a tiny FID peak (arrow). Top right, molecular structure of the EAD-active compound.

pale-brown chafer, *Phyllopertha diversa* (Coleoptera: Scarabaeidae). It is interesting that this compound was synthesized more than 40 years ago, and has been reported to have anti-inflammatory^{2,3}, analgesic and anticonvulsant effects³. We report its first isolation from a natural source.

The pale-brown chafer is a devastating turf pest in Japan, for which environmentally sound methods of control are badly needed. Evidence for the occurrence of a strong female-released sex pheromone (which accounts for the formation of a ball of males around a single female) was obtained more than a decade ago⁴. Several groups have been trying to identify this sex pheromone, but the tiny amount released by the females (at the picogram level) prevented its characterization.

When we subjected a sample of a biologically active ether extract, obtained from females collected from the field, to gas chromatography with an electroantennographic detector⁵, we observed only one active peak (see Fig. 1). The amount of the semiochemical was so small that it was almost undetected by the flame ionization detector. Spectral data suggested that the natural product would have a benzene ring fused to a cyclic diamide moiety, that is, 1,3-dimethyl-2,4-(1*H*,3*H*)-quinazolinedione.

We prepared this compound by treating benzoylneurea with sodium hydroxide and iodomethane in dimethylsulphoxide and found that it was indistinguishable from the natural product. In the field, traps baited with the synthetic alkaloid captured as many as 153 ± 52 males per trap per hour, whereas the hourly catches per trap for the control were as low as 0.4 ± 0.5 (*t*-test, 28.0865; *P* > *F*, 0.0001).

Molecules that have signal value in nature sometimes turn out to be of use to humans⁶; well-known recent additions to our therapeutic arsenal include ivermectin, cyclosporin, FK-506 and taxol*. The discovery of an insect sex pheromone with

medicinal properties is, therefore, fortuitous, but not entirely unexpected.

It has been suggested that insect-pheromone recognition is mediated by a family of G-protein-coupled receptors^{7,8}. The analgesic effects of non-steroidal anti-inflammatory drugs are primarily attributed to the inhibition of prostaglandin biosynthesis both in the peripheral and central nervous system, but other modes of action (for example, via G-protein dependent pathways) are also considered⁹. Identification of an insect pheromone with aspirin-like anti-inflammatory and analgesic properties for mammals may therefore contribute to better understanding of the interactions of this class of drugs with the nervous system.

Walter Soares Leal

Paulo H. G. Zarbin

Hubert Wojtasek

Laboratory of Chemical Prospecting,
National Institute of Sericultural and
Entomological Science,
Tsukuba, Ibaraki 305, Japan
e-mail: leal@nises.affrc.go.jp

Shigefumi Kuwahara

Laboratory of Agricultural Chemicals,
Ibaraki University, Ami-machi,
Inashiki-gun,
Ibaraki 300-03, Japan

Makoto Hasegawa

Chiba Prefectural Agricultural Experiment Station,
808 Daizenno-cho, Midori-ku,
Chiba 266, Japan

Yasuo Ueda

Ibaraki Agricultural Center,
3344 Kamikuni, Mito-shi 311-42, Japan

1. Mayer, M. S. & McLaughlin, J. R. *Handbook of Insect Pheromones and Sex Attractants* (CRC, Boca Raton, 1990).
2. Maillard, J. *et al. Chim. Ther.* **3**, 100–106 (1968).
3. Montginoul, C., Pastor, G., Vigne, C. & Giral, L. *Ann. Pharm. Fr.* **46**, 223–232 (1988).
4. Kawasaki, K. & Tamaki, Y. *Appl. Entomol. Zool.* **20**, 137–142 (1985).
5. Leal, W. S., Kuwahara, S., Ono, M. & Kubota, S. *Bioorg. Med. Chem.* **4**, 315–321 (1996).
6. Fisher, T. & Meinwald, J. *Chemical Ecology: The Chemistry of Biotic Interaction* (National Academy, Washington DC, 1995).
7. Breer, H. *Semin. Cell Biol.* **5**, 25–32 (1994).
8. Shepherd, G. M. *Neuron* **13**, 771–790 (1994).
9. Walker, J. S. *Clin. Exp. Pharmacol. Physiol.* **22**, 855–860 (1995).

*Bristol-Myers Squibb has registered Taxol as a trademark and wishes the scientific community to use the name paclitaxel.

Glucagon-like peptide-1 and satiety

On the basis of their observation that intracerebroventricular administration of glucagon-like peptide-1 (residues 7–36) amide (GLP-1) reduced food intake in rats, Turton *et al.*¹ suggest that GLP-1 is a physiological mediator of satiety. Using c-Fos immunohistochemistry as a marker of neuronal activity, they reported that GLP-1 induces c-Fos expression “exclusively” in the paraventricular hypothalamus and central amygdala, regions of the brain important in the regulation of feeding. Although consistent with the hypothesis that GLP-1 causes satiety, there are alternative interpretations for the earlier result. We report here that GLP-1, in addition to reducing food intake^{1,2}, also supports the development of a strong conditioned taste aversion, indicating that GLP-1 has aversive side-effects that could attenuate consumption and hence provide one explanation for the data of Turton and colleagues¹.

We administered GLP-1 (0.3, 1.0 or 3.0 µg) intracerebroventricularly to undeprived rats at the beginning of the dark phase of their light–dark cycle. Consistent with ref. 1, 3.0 µg was the lowest dose to produce reliable reductions in food consumption over a 2-hour period (Fig. 1a). To examine the potential of GLP-1 to produce a conditioned taste aversion, we infused saccharin directly into the oral cavity of

other rats³, followed immediately by GLP-1 administration (1.0 µg or 3.0 µg). When re-exposed to saccharin several days later, a significant number of rats rejected the taste when it was paired with 3.0 µg but not 1.0 µg of GLP-1 (Fig. 1b). Rats that received saccharin that had previously been paired with intraperitoneal injection (63 mg kg⁻¹) of the emetic agent lithium chloride displayed a conditioned taste aversion similar in magnitude to 3.0 µg GLP-1 (data not shown). Saccharine exposure followed by administration of control solutions for GLP-1 or LiCl did not cause rats to reject saccharin during testing for conditioned taste aversion, nor did exposing rats to GLP-1 or LiCl that was not previously paired with saccharin. Thus, reduction of food intake by GLP-1 is apparently accompanied by symptoms sufficient to support a conditioned taste aversion.

We have also examined the pattern of c-Fos induction after GLP-1 administration. Consistent with ref. 1, we find that c-Fos is expressed in the paraventricular nucleus and central amygdala². GLP-1 also induces strong c-Fos expression in the nucleus of the solitary tract, area postrema and pontine parabrachial nucleus, eliciting a pattern of brainstem activity similar to that elicited by LiCl³.

In summary, we find that intracerebroventricular GLP-1 generates neural activity in the brainstem, as assessed by c-Fos immunohistochemistry², strikingly similar to that of a known emetic agent³. We have also found that the lowest dose of GLP-1 to reduce food intake supports

the development of a conditioned taste aversion. These observations suggest that nonspecific, aversive symptoms may be elicited by GLP-1 and that these symptoms could account for observed reductions in food intake. It has often been cautioned that hormonal and pharmacological treatments can lead to reductions in food intake for various reasons, and that criteria for classifying satiating agents need to exclude other possible, nonspecific factors (see, for example, ref. 4). Our observations argue for a more careful consideration of GLP-1 before classifying it as a satiating agent.

**G. van Dijk, T. E. Thiele
R. J. Seeley, S. C. Woods
I. L. Bernstein**

Departments of Psychology and Medicine,
University of Washington,
Seattle,
Washington 98195, USA
e-mail: gvandijk@u.washington.edu

Bloom et al. reply — van Dijk *et al.* provide interesting and valuable new data. But exogenously administered neurotransmitters penetrate a wide area in an abnormal concentration gradient. They probably activate multiple circuits simultaneously, particularly when at grossly supraphysiological concentrations.

In contrast, we based our observations on the administration of a specific antagonist that acts only to block endogenous peptide release and thus affects only physiologically active circuits. Blocking endogenous GLP-1 release causes healthy animals to eat more. Since the publication of our paper¹, we have administered the GLP-1 antagonist exendin (residues 9–39) to rats twice daily for 10 days; the animals not only significantly increase their 24-h food intake but also gain a statistically significant amount more weight than controls. Our view therefore remains that endogenous GLP-1 is indeed involved in the regulation of satiety, irrespective of its other possible functions.

S. R. Bloom†

Division of Endocrinology and Metabolism,
Hammersmith Hospital Medical School,
London W12 0NN, UK
e-mail: sbloom@rpms.ac.uk

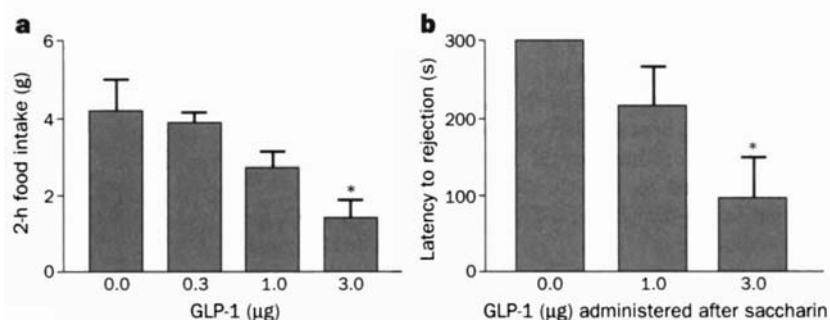


Figure 1 a, Food intake in undeprived rats 2 h after infusion of synthetic cerebrospinal fluid (sCSF; $n=6$; left bar) or GLP-1 (0.3 µg, $n=7$; 1.0 µg, $n=6$; 3.0 µg, $n=4$) just before the dark phase ($F_{3,19}=4.27$, $P<0.05$; GLP-1 versus sCSF; *, $P<0.01$). **b**, Latency to reject intraoral infused saccharine solution (2.5 ml of 0.15% over 5 min) by undeprived rats that had previously received saccharin immediately before receiving sCSF ($n=5$) or GLP-1 (1.0 µg, $n=6$; 3.0 µg, $n=5$). Rats more rapidly rejected saccharin paired with 3.0 µg GLP-1 ($F_{2,13}=5.17$, $P<0.05$; 3.0 µg GLP-1 versus sCSF; *, $P<0.05$). Male Long–Evans rats (300–350 g) were housed individually in stainless-steel cages and given free access to food and water. Cannulae were positioned and assessed as described elsewhere^{2,3}. Intracerebroventricular drugs were dosed in a 3.5-µl volume. Immediately after saccharine exposure, rats were given either GLP-1 (1.0 µg or 3.0 µg; paired GLP-1 group), LiCl (63 mg kg⁻¹, 0.15 M; paired LiCl group) or the proper controls of both treatments (sCSF; unpaired GLP-1 or isotonic saline; unpaired LiCl). Several days later and without exposure to saccharin, rats in paired groups received control treatments whereas unpaired rats received GLP-1 (3.0 µg) or LiCl. Thus all rats had equal exposure to drugs and to saccharin, either on the same day (paired groups) or on separate days (unpaired). Finally, latency to reject intraoral saccharin was assessed (active fluid expulsion and/or passive dripping from the oral cavity).

1. Turton, M. D. *et al.* *Nature* **379**, 69–72 (1996).
2. van Dijk, G. *et al.* *Am. J. Physiol.* **271**, R1096–R1100 (1996).
3. Swank, M. W., Schafe, G. E. & Bernstein, I. L. *Brain Res.* **673**, 251–261 (1995).
4. Deutsch, J. A. in *Handbook of Behavioral Neurobiology* Vol. 10: *Neurobiology of Food and Fluid Intake* (ed. Stricker, E. M.) 151–182 (Plenum, New York, 1990).

†On behalf of the coauthors of ref. 1.

scientific correspondence is intended to provide a forum in which readers may raise points of a scientific character. Priority will be given to letters of fewer than 500 words and 10 references.

How to succeed in science

The Scientific 100: A Ranking of the Most Influential Scientists, Past and Present

by John Simmons

Citadel: 1996. Pp. 504. \$29.95

W. F. Bynum

We have the *Financial Times* 100 top shares and *Fortune Magazine's* 100 wealthiest people, the 100 best-dressed women and the 100 most eligible bachelors, the 100 best wines and the 100 greatest films. More than a century ago, John Lubbock published *The Hundred Best Books* (his was not one of them). Why not *The Scientific 100*?

Even if scientific influence might be more difficult to measure than market value or personal wealth, it ought to be as easy to judge as dress sense or personal desirability. And comparing a naturalist with a theoretical physicist, or a chemist with an anthropologist, should pose no more problems than ranking a claret alongside a pinot noir. There is even a century-old precedent for scientific rankings — James McKeen Cattell based his league table on the number of column inches given to scientists in encyclopaedias.

John Simmons is more egalitarian: his biographical sketches of his ranked 100 best scientists are each roughly the same length, four or five pages. Wilhelm Wundt (number 99) gets about as much space as Isaac Newton (1), and Edward O. Wilson (83) as much as Niels Bohr (3).

Further, Simmons admits that a rough layering rather than a precise ranking might be more appropriate: no one is likely to argue that Ernst Haeckel (90) ought to go above Charles Darwin (4), but they might rank him higher than Wilson, or object to his (or Wilson's) inclusion in the first place. Archimedes is placed last, as a founder of the whole scientific tradition, but he could have gone first; however, the whole list could not be reversed without perversity. Simmons's enterprise is arbitrary, but not completely so. Scientific influence (or worthiness) is difficult to define, but not an illusion; a quartile, if not an absolute rank, makes intuitive sense.

These caveats aside, and with a bottle of first-quartile claret to hand, Simmons's volume becomes enjoyable. It is well written, has a reasonable standard of accuracy and a defensible selection. A disarming appendix discusses "inexcusable omissions, honorable mentions and also-rans". Simmons singles out 50 of these including Aristotle, Hippocrates, Descartes, Robert Boyle, Jons Berzelius and Wolfgang Pauli.

He does not apologize for excluding Leibniz, Thomas Young, Alexander von Humboldt, Peter Medawar or Karl Pearson. Ivan Pavlov probably has a better case for inclusion

than B. F. Skinner (98), Ramón y Cajal than Emil Kraepelin (92), John Enders or Max Theiler than Jonas Salk (91), Howard Florey than Alexander Fleming (97), Georges Cuvier than George Gaylord Simpson (78). Franz Boas (14) and Buffon (23) are flattered by their rankings, and Trofim Lysenko (93) earns his place only as a cautionary tale, here sensibly directed not against Soviet ideology but American creationist science. With two anthropologists, and Noam Chomsky (71) representing linguistics, there is a broad definition of 'science'. Inventors and technologists are excluded.

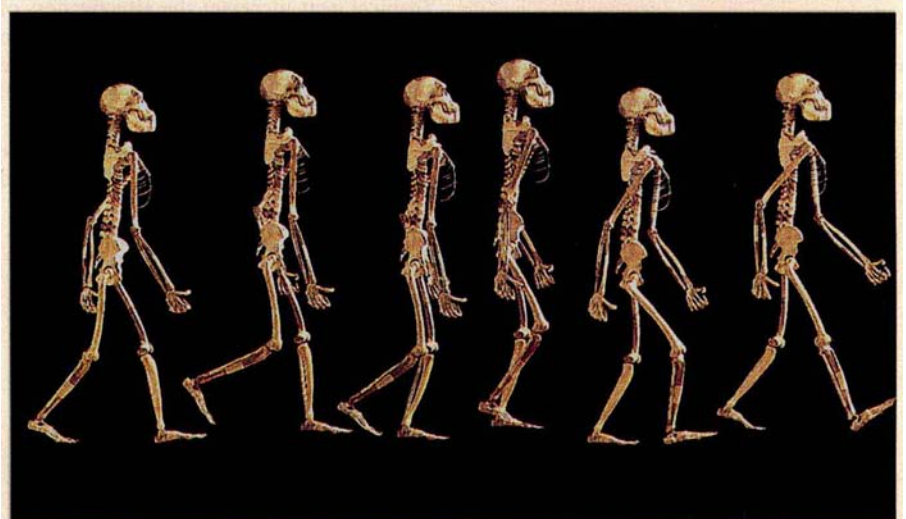
On the assumption that Simmons's ranking is inevitably subjective but more or less sound, what lessons could a young scientist learn who wanted to earn a place in a future 100? First, be male, white and middle class. Simmons has identified only three women (Marie Curie, 26, Lynn Margulis, 80, and Gertrude Elion, 85) and no Asians or Africans. Genteel poverty is perfectly acceptable, but only Faraday (11) and Lysenko could boast of real proletarian origins.

Second, a German birth helps. A quarter of his top 100 were born there, although a quarter of those ended up working in the United States. The Nazi-inspired emigration of scientific talent is prominent in Simmons's bio-

graphical sketches. The United Kingdom (with 18 scientists listed) ranks second to Germany, with the United States (17) just behind. Of those born in the United States, several had parents who were recent immigrants (Gell-Mann, 45, Glashow, 48, Oppenheimer, 87). The French, more comfortable at home and headed by Pasteur (5), come fourth, with nine listings, followed closely by the Austrians with seven (Freud, 6, tops these). The Austrian émigrés were more selective in their ports of call, including the United Kingdom (Freud), Ireland (Schrödinger, 18) and the United States (Landsteiner, 81, but before the Nazi period).

The current status of American science is highlighted by the fact that all but three of Simmons's living selections reside there (the exceptions are Hawking, 54, Sanger, 72, and Levi-Strauss, 79). Some crumb of American comfort can be derived from the fact that seven of the remaining 11 living scientists were born in the United States, even though the highest-ranked living scientist is British, Francis Crick (33).

A young physicist hoping to reach the top could be advised to head for California (home to most nuclear physicists), whereas life scientists can afford to stay on the East Coast. If a Nobel prize is a priority, avoid evolutionary biology (neither Mayr, 65, Dobzhansky, 67,



Walk-on part in human evolution

Australopithecus afarensis lived between 3 and 4 million years ago and is thought to be the earliest known bipedal ancestor of the human species. The first specimen to be found, the partial skeleton known as Lucy, achieved fame and was dubbed "the woman who shook up man's family tree". Lucy's discoverer, the palaeoanthropologist Donald Johanson, has collaborated with the science writer Blake Edgar to produce *From Lucy to Language*

(Simon and Schuster/Weidenfeld and Nicolson, \$50, £25), an encyclopaedic overview of human evolution. The book, to be reviewed in a future issue of *Nature*, includes more than 200 colour plates of the most valuable hominid fossils, artefacts and prehistoric art, many reproduced at actual size. Most of the photos were taken by David Brill. The picture shows a plaster reconstruction of *afarensis*.

nor Simpson, 78, has received the award). If influence with the Pentagon is a goal, be born in Hungary and emigrate (von Neumann, 51, and Teller, 88). Do not worry about suicide (only two, Boltzmann, 24, and Fischer, 46), or an otherwise violent death (Lavoisier, 8, and Archimedes). Expect a reasonable lifespan. Only Clerk-Maxwell (12) and possibly Euclid (59) died before the age of 50, and the living scientists already have an average age of 74. Ernst Mayr, born in 1904, energetically heads this group.

If this is beginning to sound like a parlour-game, that is no surprise: Simmons's methods do not bear close scrutiny. Nevertheless, he has brought off his assignment with a good deal of aplomb, and his volume conveys both the excitement and the human dimensions of science. It may not satisfy the purists, but it offers portraits of 100 remarkable individuals who have changed the way we think about the world and ourselves. □

W. F. Bynum is at the Wellcome Institute for the History of Medicine, 183 Euston Road, London NW1 2BE, UK.

Walking wounded

The Killing Factory: The Top Secret World of Germ and Chemical Warfare

by John Parker

Smith Gryphon: 1996. Pp. 230. £16.99

Alistair Hay

Preparation for chemical and biological warfare is still a very secretive affair. Anyone who attempts to peer behind the veil that governments use to hide their activities in this area will find it firmly drawn. John Park-

er's book is claimed to investigate "the top secret world of germ and chemical warfare". The author does describe much that was previously secret, but it is a history that has been revealed by others over the years.

With all the information now available, there is ample subject matter for Parker's pen. As an experienced journalist he has an eye for a good story, and his text races along, noting most of the important milestones. In addition, Parker includes valuable testimonies from individuals who experienced the effects of chemical agents.

These testimonies provide revealing, but heartbreaking, reminders that the victims are not just the poor bloody infantry, or civilians caught downwind of an exploding bomb, but also those who worked with the chemical agents or volunteered to take part in tests. Over the years thousands of servicemen have voluntarily participated in programmes either to investigate the properties of chemical warfare agents or to develop antidotes.

During the Second World War, British, American and Australian servicemen took part in some of the most testing of these programmes in northern Queensland, to investigate the incapacitating properties of mustard gas. This chemical agent poisons through inhalation or skin contact. Protective clothing and a respirator are required for full protection, but there are no antidotes.

The author notes that in 1943 there was concern that Japanese forces might use mustard gas in the Pacific theatre of war. Tropical heat increases the effectiveness of the gas, and military planners needed to know whether exposure to it would disable soldiers totally or whether some might still be able to fight.

Most of those who volunteered to take part in the experiments wore respirators and

were exposed to the gas in purpose-built chambers. Twenty-four hours after exposure, significant blistering occurred on the skin, and days later the blistered area had become painfully ulcerous. The body areas most severely affected were those that became hot and sweaty — the groin, armpits and wherever clothing was tightly pressed against the skin. On the days following their exposure to the gas, the volunteers were sent on route marches or over obstacle courses until their disability was such that they could barely walk. Some of these walking wounded were still able to man defensive positions, whereas others required hospital treatment.

Ethical approval for the chemical warfare trials carried out in Australia in 1943–45 would not be forthcoming today. Parker says that at the time, and even today, chemical warfare experts, as well as volunteers, felt that the programme provided valuable information. But it is to the shame of the governments concerned that for decades they denied that these trials ever took place. Eventually, and only some six years ago, the Australian government agreed to compensate survivors for related injuries.

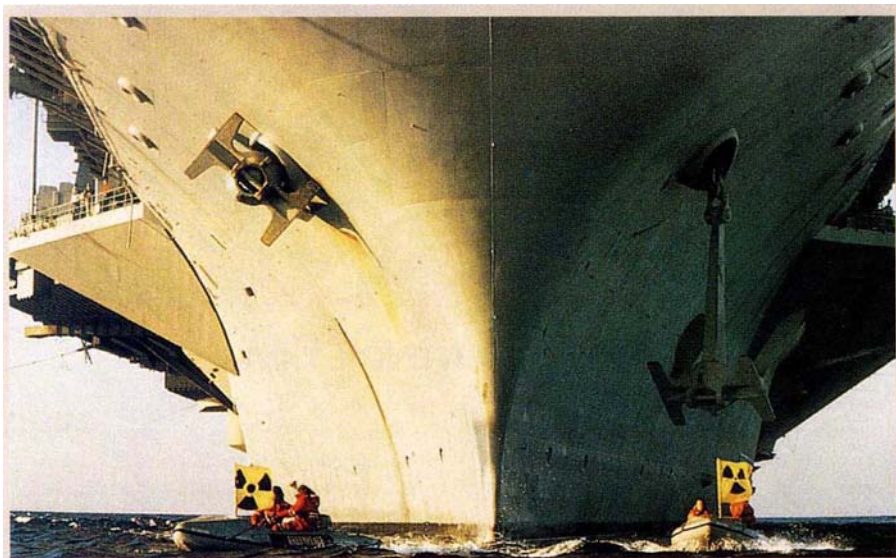
Volunteers who took part in trials in the United Kingdom are less fortunate. The UK government denies that they have any health problems. How it can arrive at this conclusion without conducting any systematic investigation on some of the thousands who volunteered has yet to be explained.

The injustice of this position leaps from the pages of *The Killing Factory* as Parker recounts case after case of volunteers who knew little about the consequences of what they had agreed to undertake. The author is also outraged about the treatment of servicemen and women suffering from what has come to be known as 'Gulf War syndrome'. Many of the troops who fought against Iraq in 1991 have since complained of ill health.

The cause of their ill health is unclear. Some claim that it is due to exposure to low concentrations of the nerve agent Sarin, which was released in the bombing of Iraqi munitions dumps. Others blame either the vaccines they were given against biological agents, or the prophylactic against nerve gas which they took in tablet form three times a day.

Until recently the UK government, unlike the US administration, had refused to provide funding for any epidemiological investigations. Consistently sceptical about the existence of any syndrome, the UK government has just agreed to fund two large-scale investigations. These will assess the extent of ill health in Gulf War veterans, and investigate whether service in the Gulf has had any adverse effect on reproduction, including rates of miscarriages and birth deformities.

Stories about some milestones in the his-



Greenpeace campaigners challenge the military might of the USS Eisenhower in a demonstration against nuclear weapons. From Greenpeace

Witness: 25 Years on the Environmental Front Line by Kieran Mulvaney and Mark Warford (Andre Deutsch, £20).

tory of chemical and biological warfare that the book refers to cannot be repeated often enough. Parker mentions the fatal biological warfare experiments that Japanese scientists carried out on 3,000 people in Manchuria during the Second World War. The research was cruel in the extreme, yet many of those involved escaped prosecution in a deal with the US government in which the scientists traded their research findings for freedom.

Although the author refers to international treaties prohibiting chemical and biological warfare, he inclines to the view that these will not stop terrorists or "a few unstable governments and dictators". So the history he recounts is but a preparation for a future in which we will still have such warfare. This is a distinctly pessimistic outlook which tends to be fostered by the intelligence community, ever anxious to increase its sphere of influence.

We must all hope that the diplomats have the final say and that their treaties keep the peace. □

Alistair Hay is in the Department of Toxicology, University of Leeds, Leeds LS2 9JT, UK.

Unified views

In Search of the Ultimate Building Blocks

by Gerard 't Hooft

Cambridge University Press: 1996. Pp. 191
£27.95, \$39.95 (hbk); £9.95, \$14.95 (pbk)

Frank Wilczek

Gerard 't Hooft is a distinguished theoretical physicist whose ideas and opinions should be of great interest to anyone curious about the development of physics in the late twentieth century. In this nonmathematical but demanding little book, he gives his personal perspective on the areas of physics with which he has been concerned — chiefly the theory of elementary particles and the quantum theory of gravity.

The book is in three parts. The first defines what elementary particle physics is and describes its status in 1970, when 't Hooft learned it as a graduate student. Authors attempting to convey this extensive and richly structured material, profoundly alien to our everyday experience but having a tight logic and austere beauty, in a few pages at semipopular level faces difficult choices. They must compromise between explaining one or two topics at length, or many telegraphically. 't Hooft chooses the latter course, and, as a result, phrases such as "it can be shown that" or "it turns out that" appear frequently. With this limitation, the discussion is accurate and perceptive, as one would expect from such an eminent authority.

The second part might be called narra-

In retrospect chosen by John C. Marshall

The Physical Basis of Mind

by George Henry Lewes
(1877)

Consider a book in four sections. Part 1 begins with cell biology, moves to the definition of an organism and concludes with a detailed discussion of evolutionary theory. Part 2 outlines the evolution of the nervous system, the principles of developmental embryology and the physiology of nervous activity. Part 3 uses this survey of neuroscience to attack dualism, idealism, mind-brain identity theory and materialist reductionism. Mind, it is argued, is necessarily embodied, but only confusion can arise from supposing that "my Body" is an expression of the same type as "my House". An account of consciousness is proposed that stresses the role of selective attention, and the notion that the unconscious is the *latent* conscious is developed. In Part 4, distinctions are drawn between mechanism and organism. Sensibility can be manifest only in organic substances, and sensation is manifest in the brain and the spinal cord. The obvious objections to the latter claim are countered by subtle 'disconnectionist' arguments. The reactions of parts of an organism "separated from the rest" are not "typical of their reactions when forming constituents of the normal organism".

What is this? The latest opus from Patricia Churchland, Gerald Edelman, John Searle or Francis Crick? No, the book is *The Physical Basis of Mind* (1877) by George Henry Lewes. But for any neuroscientist now opening this work for the first time, the sense of partial familiarity is palpable. Over the past decade or so, it has become conventional to refer to the 'mind/brain' as if the slash *per se* resolved the problem of dualism. Twelve decades earlier, Lewes employed the hyphen: "We know ourselves as Body-Mind; we do not know ourselves as Body and Mind, if by that be meant two coexistent independent Existents." But unlike some of today's writers for whom the study of consciousness is a form of topless bungee jumping (without the rubber rope), Lewes also deploys experimental data and coherent argument.

Gilbert Ryle's notion of a category-mistake is lucidly anticipated: "To ask if Matter could think, or Mind move Matter" is, Lewes asserts, "a confusion of symbols equivalent to speaking of a yard of Hope, and a ton of Terror". And the pernicious results of category-mistakes can lead the neuroscientist to write such nonsense as: "A sensory impression is transmitted as a wave of motion to the brain, and there being transformed into a state of consciousness, is again reflected as a motor impulse."

Lewes's way out is to develop a

sophisticated version of dual-aspect theory.

The laws of mechanics cannot be violated:

"the abstract laws of movement for an organic body are not different from the abstract laws of movement for an inorganic body". But neither reductionism nor constructivism suffices: qualitative knowledge will not "translate... into physiological terms". Similarly, "nor can the laws of Mind be deduced from physiological processes unless supplemented by and interpreted by psychical conditions individual and social".

Indeed, Lewes always thought that the artist would provide insights into the life of the spirit complementary to those of the natural scientist. His great biography *The Life and Works of Goethe, with Sketches of His Age and Contemporaries* (1855) successfully unites the author of Faust with the aggressively anti-Newtonian colour theorist. Lewes himself had little talent for creative writing, but shortly after finishing his book *Comte's Philosophy of the Sciences* (1853) he eloped to the Continent with Marian Evans (a.k.a. George Eliot). The 24 years they spent together produced *Middlemarch* and *Daniel Deronda* as well as *The History of Philosophy from Thales to Comte* and the five volumes of *Problems of Life and Mind* (*The Physical Basis of Mind* is the fourth). Their relationship was such that neither body of work could have existed without the mutual support (emotional and intellectual) they provided each other.

After her lover's death in 1878, Marian gave more than £5,000 of her royalties to establishing the George Henry Lewes Studentship in Physiology. The first holder was C. S. Roy, who eventually became professor of pathology at Cambridge; a slightly later recipient was Charles Sherrington. Lewes had been fascinated by the possibility of functional brain imaging *in vivo*. In *The Physical Basis of Mind* he outlines the initial steps taken towards showing how sensation and ideation may produce local changes in blood flow and raise brain temperature. He would have been pleased that 'his' students, Roy and Sherrington (1890), were the first to describe "an automatic mechanism by which the blood supply of any part of the cerebral tissue is varied in accordance with the activity of the chemical changes which underlie the functional action of that part".

John C. Marshall is in the Neuropsychology Unit, University Department of Clinical Neurology, Radcliffe Infirmary, Oxford OX2 6HE, UK.

"If any look to him for comfort they will find that, promising them bread, he gives them a stone — the same stone that has already set their teeth on edge." From the original review in *Nature* 16, 261 (1877).

tive-historical. The author gives a description of the remarkable period 1970–76, when the foundations were laid of the Standard Model of fundamental particles and their interaction, and weaves this around a first-person account of the early part of his own career. Because his work was central to much of the progress at the time, this approach is not unnatural, and on occasion offers a unique perspective. In particular, 't Hooft's account of his encounters with his adviser and collaborator Martinus Veltman, and with Ben Lee, which led up to his historic proof of the renormalizability of non-abelian gauge theories, is a fresh and interesting contribution to intellectual history.

But there is danger in such a personalized approach, regarded as history, when the author attempts to assess his own contributions. For example, I believe, from my own experience as a participant, that he seriously understates the conceptual and technical achievements of others in discovering and recognizing the significance of asymptotic freedom, and in using it to construct a specific and testable theory of the strong interaction. It was quite a bold step, in my opinion, to propose a theory whose building-

blocks were entirely unobserved at the time; and it was credible only in the light of difficult, specific calculations that offered an explanation of surprising experimental facts (Bjorken scaling, observed at the Stanford Linear Accelerator Center in California). The achievement of the modern theory of the strong interaction belongs to those who made these discoveries and took these steps.

Similarly, it is more than a little jarring, for anyone who lived through the period, to find the main concepts of charmonium theory expounded in reconstructed dialogues between the author and bystanders, with no mention of Appelquist and Politzer's fundamental paper, or of the earlier work of Lee, Gaillard and Rosner.

The third part of the book, in which 't Hooft discusses attempts by himself and others to go beyond the Standard Model, provides the most fun. He loosens his pen and gives his frank opinions on what are still very much live issues. He expresses some of his reservations about string theory in a memorable image: "Actually, I would not even be prepared to call string theory a 'theory' but rather a 'model', or not even that: just a hunch.... Imagine that I give you a chair

while explaining that the legs are still missing, and that the seat, back, and armrest will perhaps be delivered soon; whatever I did give you, can I still call it a chair?"

In a more serious vein, he expresses his conviction that string theory may provide valuable insight, but that a truly fundamental theory of physics will ultimately be based on entirely different concepts. Physicists may be startled to learn that 't Hooft, like Einstein, de Broglie and several other heroes of quantum physics, expresses considerable sympathy for the idea that quantum mechanics will eventually be supplanted by some form of 'hidden variables' theory. He is also fascinated by the vision of Wheeler, Fredkin and others that the ultimate laws might be of some elementary logical-combinatorial form, along the lines of a universal (in both senses!) cellular automaton.

Despite its serious flaws, one finds on virtually every page of this book sharp statements and novel formulations that show the workings of a first-rate, confident and original mind. It deserves attention. □

Frank Wilczek is at the School of Natural Sciences, Institute for Advanced Study, Princeton, New Jersey 08540, USA.

At a glance

Excellent ★★★★★ Good ★★★★ Fair ★★★ Poor ★

Brain Mapping: The Methods

by Arthur W. Toga and John C. Mazziotta
Academic: 1996
Pp. 471. \$145

Investigation of the functional architecture of the human brain using modern noninvasive imaging techniques is a rapidly expanding area of research. A proper knowledge of methodology is needed to appreciate the burgeoning literature in the field.

Now we have an excellent catalogue of the main techniques. The editors have assembled an impressive group of experts, all widely known in their field, who contribute an outstanding set of chapters.

The editorial hand has been firm and judicious, placing each contribution in context so that the book develops logically. Coverage ranges from recording intrinsic optical signals, through tomographical mapping techniques and electrically and magnetically based measurements, to issues of experimental design and statistics relevant to image analysis.

I cannot recommend this volume highly enough for both experts and those starting out in the field. It is also beautifully illustrated and produced.

R. S. J. Frackowiak Institute of Neurology, London, UK.

Range	★★★★
Depth	★★★★
Accuracy	★★★★
Up-to-dateness	★★★★
Accessibility	★★★★
Style	★★★★

Triple-Helical Nucleic Acids

by Valery N. Soyfer and Vladimir N. Potaman
Springer: 1996. Pp. 360.
\$79, £60

A thriving new area of research in nucleic acids focuses on triple helical structures in DNA. These structures can be thought of as generated by adding a third strand to duplex DNA through the formation of new base-specific interactions in the major groove. The potential application of triplex structures as targeted therapeutic agents now attracts academic and commercial interest.

The authors review what we know about most aspects of these novel structures, covering structural, molecular recognition and physico-chemical aspects in detail. They also describe the methodology used, although I was surprised at the lack of nuclear-magnetic-resonance evidence throughout. But perhaps my greatest reservation, both about the book and the field, came on reading the penultimate chapter on *in vivo* significance. The authors state in the preface that the "importance [of triplex structures] in many key processes of life is now obvious"; well, not to me it isn't. Read the book to learn the structural aspects of triplexes, but take the biology with caution.

David Lilley Department of Biochemistry, University of Dundee, UK.

Range	★★★
Depth	★★
Accuracy	★★
Up-to-dateness	★★
Accessibility	★★★★
Style	★★★

Ion Channels: Molecules in Action

by David J. Aidley and Peter R. Stanfield
Cambridge University Press: 1996. Pp. 307.
£50, \$74.95 (hbk); £17.95, \$29.95 (pbk)

The title of this textbook conveys the sense not only of the active switching of these macromolecular machines, but also of the extraordinarily dynamic nature of the field of ion-channel studies.

Since channel proteins were first cloned 15 years ago, there has been a barrage of new molecular and functional information about ion channels, and in the 1990s more and more channels have been linked to inherited human diseases.

Here two eminent physiologists review basic principles and historical studies, as well as emphasizing the latest molecular advances. They begin at a basic level suited to undergraduate biologists with little background in physics, but work towards a fairly sophisticated understanding of current research.

The survey of newly cloned channels, many of which are outside the historical neuronal channel families, should give newcomers a feeling for how much remains to be explored.

Gary Yellen Department of Neurobiology, Massachusetts General Hospital and Harvard Medical School, USA.

Range	★★★★
Depth	★★★
Accuracy	★★★
Up-to-dateness	★★★★
Accessibility	★★★★
Style	★★★

Mantle geochemistry: the message from oceanic volcanism

A. W. Hofmann

Basaltic volcanism 'samples' the Earth's mantle to great depths, because solid-state convection transports deep material into the (shallow) melting region. The isotopic and trace-element chemistry of these basalts shows that the mantle contains several isotopically and chemically distinct components, which reflect its global evolution. This evolution is characterized by upper-mantle depletion of many trace elements, possible replenishment from the deeper, less depleted mantle, and the recycling of oceanic crust and lithosphere, but of only small amounts of continental material.

Geochemists began to study the composition and evolution of the Earth's mantle in the 1960s, when it became clear that plate tectonics is driven by solid-state mantle convection, which carries deep-mantle material upwards until it begins to melt (at about 30–100 km depth). This melt rises to the surface where it extrudes as basaltic lava and delivers a chemical 'message' from the mantle to the geochemist.

When a mantle region melts, it loses **incompatible** (see Box 1 for explanation of professional jargon) trace elements, such as uranium, thorium and potassium, to the melt, which transports them to the oceanic or continental crust. The continental crust now contains about half of the Earth's total inventory of Th and U, leaving at least the upper mantle correspondingly **depleted**. This depleted region is sampled globally by basalts along the 60,000-km-long mid-ocean-ridge system. Other types of volcanism are fed by less depleted or even **enriched** regions in the mantle. A mantle region can 'remember' its depletion or enrichment history, because the accumulation of daughter products of radioactive decay records the changes in parent-to-daughter concentration ratios which occur during melt removal (or addition). Because of this, isotopic measurements have become the major (though not the only) tool of the mantle geochemist. This allows us to 'map' the geochemical structure of the Earth's interior by interpreting the isotopic composition of mantle-derived basalts. Thus isotope geochemistry, having established the chronology of the Earth's accretion and continent formation, is now beginning to decipher its internal differentiation history as well.

I review here the main geochemical evidence bearing on mantle structure and processes, and some of the models based on it. I conclude that much of the chemical heterogeneity evident in oceanic basalts is caused by **recycling** of oceanic, and to a much lesser extent continental, crustal material, which resurfaces predominantly in the 'mantle plumes' that create volcanic islands. Whether these plumes rise from the base of the upper mantle or the base of the lower mantle remains a hotly debated issue. Earlier reviews of this and related topics may be found in refs 1–9.

Overview of mantle models

In the 1970s, geochemists developed the idea of a chemically layered mantle, with an upper depleted and a lower undepleted (also called **primitive**) layer, because it was found that there is a fundamental geochemical difference between mid-ocean-ridge basalts (**MORB**), which come from mantle regions that are depleted in all incompatible elements, and ocean island basalts (**OIB**), which are less depleted or even enriched. Ocean islands often form long lines of volcanoes of progressively increasing ages, which are created by an apparently stationary heat source ('**hotspot**') under a moving plate¹⁰. Morgan¹¹ proposed that hotspots are created by narrow 'plumes' rising from the deep mantle. The chemical and isotopic

differences between MORB and OIB appeared to indicate that the former come from an upper, depleted layer, whereas the latter rise from a deeper, undepleted layer (see, for example, ref. 12). This is the geochemical 'standard model' shown in Fig. 1a. Apparent support came from geophysical evidence showing a sharp increase in the velocity of seismic waves at a depth of 660 km (the **660-km seismic discontinuity**). This indicated that the mantle consists of two layers, which differ in either chemical or mineralogical composition or both.

This standard model is now untenable because OIB are highly variable in composition, so that they could not be derived from a chemically uniform, primitive layer of the mantle. This has led to a proliferation of speculative models about the origin and distribution of the different types of material in the mantle. Some of these have retained the concept of a two-layer mantle (Fig. 1b): the two convecting layers are kept separate by an intrinsic compositional density contrast or by a negative pressure–temperature slope of an isochemical phase transition at 660 km (ref. 13). Plumes originate from the base of the upper layer but may entrain material 'leaking' from the lower layer^{14,15}. Other models are based on 'whole-mantle' convection (see, for example, ref. 16 and Fig. 1c). In this class of models, **subducted lithosphere** penetrates the 660-km boundary, and all plumes originate from the core–mantle boundary. Recently, hybrid models have been developed in which the two, normally separately convecting layers are intermittently perturbed when sinking pieces of subducted lithosphere and rising plumes occasionally penetrate the 660-km boundary (Fig. 1d; refs 5, 17, 18). It will be seen further below that geochemical mass-balance considerations strongly favour some type of layered convection (models in Fig. 1b or d).

Geological background

The following geological observations and inferences derived from them provide a framework for interpreting the geochemical messages from the mantle:

(1) Mid-ocean ridges migrate laterally in a pattern controlled largely by the surface geometry of the plates. Therefore, they should normally 'sample' the uppermost mantle (that is, the **asthenosphere**), which rises passively under the spreading ridge. Occasionally, a migrating ridge intersects a plume, which rises from the deeper mantle. When this happens, such as on Iceland, oceanic crust of anomalous chemistry and thickness is produced.

(2) Mantle plumes are thought to generate about 20–40 stationary 'hotspots', which are also used as a reference frame to determine absolute plate motions. Plumes probably originate from boundary layers in the mantle^{7,19,20}, which may be located either above the 660-km seismic discontinuity or above the core–mantle boundary (at 2,900 km). In either case, heating from below lowers the density until the layer becomes unstable and forms a rising column or 'plume'.

Box 1 Geochemical and geophysical jargon

Asthenosphere. Relatively soft, hot mantle region (at or near the melting point) underlying the colder, harder **lithosphere**. Its thickness is of the order of 100 km but is poorly defined because stiffness increases gradually with depth.

Depleted (enriched) reservoir. Region in the mantle that is depleted (enriched) in **incompatible elements** relative to a primitive reservoir.

Hotspot. Locus of volcanism that remains nearly stationary relative to the moving lithospheric plates. Classic hotspots form long chains of volcanoes (such as the Hawaiian-Emperor chain) which become progressively older as a function of distance from the presently active volcanism.

Incompatible element. A chemical element that is excluded from solid minerals of the upper mantle and therefore preferentially enters an available melt phase. Reasons for incompatibility include large ionic radius and high ionic charge, both of which cause misfits in the existing crystal structures of mantle minerals. Incompatible elements have, by definition, low **partition coefficients**, usually $D \leq 0.1$.

Lithosphere. Stiff, cold layer comprising the uppermost mantle and overlying crust. The lithosphere is about 100 km thick in oceanic regions and 100–400 km in continental regions. The moving plates created by sea-floor spreading consist of lithosphere.

Mantle plume. Solid-state, narrow upwelling current in the mantle with a diameter of the order of 100 km and originating from a hot, low-density boundary layer located either above the seismic discontinuity at 660 km depth or near the core-mantle boundary at 2,900 km depth.

Metasomatism. Process of changing the bulk chemical composition of a rock, usually by infiltration of an aqueous or carbonaceous fluid or melt.

MORB. Mid-ocean-ridge basalt. MORB forms the upper part of the oceanic crust.

OIB. Ocean island basalt, restricted to islands that are not related to **subduction**. Typical OIB occur on Hawaii, Iceland and the Polynesian islands, but not in so-called island arcs such as the Aleutians, Marianas or Japan.

Partition coefficient. Description of the equilibrium distribution of a trace element between different phases. According to Henry's law, the partition coefficient is independent of the absolute concentration of the trace element. For solid-melt systems, the coefficient is defined as $D = C^{\text{solid}}/C^{\text{melt}}$, where C is the concentration of the trace element in question.

Primitive mantle (sometimes called 'bulk silicate Earth'). Silicate portion of the Earth as it existed after separation of the core but before it was differentiated into crust and present-day mantle. A **primitive reservoir** is any region of the mantle that has retained this composition. The composition of the primitive mantle is inferred from the composition of the stony meteorites known as chondrites^{124,125}.

Recycling. Introduction of (oceanic or continental) crust or lithosphere into the mantle by **subduction** or delamination, storage in the mantle, and reappearance in volcanism.

Sea-floor spreading. Creation of new oceanic lithosphere by solid-mantle upwelling, consequent partial melting, extrusion and cooling of **MORB** at a mid-ocean ridge, and lateral movement of the new lithosphere away from the ridge.

Subduction. Destruction of **lithosphere** by sliding back into the mantle beneath island arcs, often at an angle of 45° and reaching depths of at least 600 km.

660-km seismic discontinuity. A narrow zone of sharp increase in seismic wave velocity at a depth of 660 km. It may be caused by a mineralogical phase change of upper-mantle silicates to denser structures, or by a sharp change in major-element composition.

(3) Not all intra-plate oceanic volcanoes originate from plumes. Thousands²¹ of small, isolated volcanic seamounts litter the oceanic crust and are probably formed by local melting anomalies. In addition, there are 'hot lines' of volcanoes that show no age progression but erupt roughly simultaneously along the entire extent of the line (for example, the Cameroon Line volcanoes²²). Most of these geochemical and melting anomalies probably come from the upper mantle.

(4) Continental flood basalts and oceanic plateaus are, according to some authors²³, formed by the surfacing of a large, mushroom-like head of a 'starting' plume, which has entrained large amounts of mantle material on the way from its origin to the surface. Often, a progressively 'younging' line of volcanoes connects an old flood basalt province (for example, the Deccan Traps of India) with a currently active hotspot (for example, Reunion Island). This is interpreted as the surface expression of the much narrower conduit or 'stem' of the plume, which may be active for over a hundred million years.

(5) **Sea-floor spreading** generates large chemical heterogeneities by extracting melt from the mantle, thereby forming a basaltic crust and a refractory residue. **Subduction** reinjects these heterogeneities back into the mantle, where they are gradually rehomogenized by convection²⁴. Subduction also causes melting, resulting in volcanic 'island arcs', which are ultimately accreted to the continental crust.

The geochemical evidence

To characterize the mantle source regions of basaltic lavas, the geochemist uses a variety of geochemical tracers. Such tracers are either isotope abundance ratios of daughter elements of radioactive nuclides, or concentration ratios of incompatible trace elements. The melt 'copies' these tracer ratios from the source and delivers them to the surface with little or no distortion. In spite of occasional assertions to the contrary, the geochemical tracing approach has withstood the test of time quite well (see, for example refs 1 and 25) and will be used here without apologies. Table 1 shows the radioactive decay systems used in this Review, their half-lives, and (radiogenic to non-radiogenic) isotope ratios of the daughter elements, which will vary as a function of age and parent/daughter element ratio (discussed below). I focus here purely on oceanic basalts (MORB and OIB) rather than mantle xenoliths or volcanic rocks from island arcs or continental regions, because oceanic basalts represent relatively large volumes of mantle and carry the smallest risk of being contaminated during magma transport through the crust to the surface.

Elemental chemistry of MORB, OIB and continental crust.

General chemical characteristics of average continental crust, average oceanic crust (= MORB), and selected OIB are shown in Fig. 2. Concentrations are normalized to those of the **primitive mantle** and plotted in the order of increasing compatibility, which, except for the 'anomalous' elements Nb, Pb and Ti (see below), coincides with the order of descending abundances in the continental crust. All types of crustal rock (MORB, OIB and continental crust) are enriched in incompatible elements (relative to the primitive mantle) because they originate as melts, which concentrate incompatible elements from the mantle. However, in the oceanic crust (= MORB), the most highly incompatible elements (for example, Rb, Ba and Th) are actually relatively depleted in comparison with the moderately incompatible elements (for example Sm and Hf), because this crust is formed by melting of a mantle region that has previously been depleted by extraction of the continental crust⁴. This effect also explains the mutually complementary positive and negative anomalies of Nb and Pb.

Ocean island basalts are much more enriched in incompatible elements than MORB, and they show a large variation in composition. This variability will be especially obvious in the isotopic compositions discussed below. Here I show average concentration patterns for basalts from Mauna Loa (Hawaii) and two of the

isotopically 'extreme' types of mantle plume (see below and Figs 3–6), namely Tubuai, Austral Islands (HIMU) and Tristan–Inaccessible Island (EM-1). (See Isotope taxonomy section, below, for explanation of HIMU, EM-1 and EM-2). Their greater enrichments in incompatible elements cause them superficially to resemble continental crust, but their Nb and Pb anomalies are similar to those of the oceanic crust (= MORB) and opposite to those of the continental crust. Thus, except for their grossly different relative levels of enrichment (which can be explained by differences in the melting process), all these basalt types have fundamentally similar trace-element patterns. In contrast, the trace-element patterns of EM-2 basalts (not shown for reasons of clarity) differ from those of other OIB and resemble that of continental crust (see section on trace-element ratios).

Isotope chemistry. The radioactive parent nuclides listed in Table 1 are all very long-lived, so the daughter isotope ratios reflect the long-term history of the basalt-source reservoirs in the mantle. Figures 3–6 show Sr, Nd and Pb isotope data for over 1,100 samples of MORB and OIB.

Strontium, neodymium and hafnium. Isotope data for Sr and Nd in MORB, OIB and continental crust are summarized in Fig. 3. The MORB points form a relatively tight cluster that defines the upper left-hand corner of the array. In general, $^{143}\text{Nd}/^{144}\text{Nd}$ correlates negatively with $^{87}\text{Sr}/^{86}\text{Sr}$, and positively with $^{176}\text{Hf}/^{177}\text{Hf}$ (not shown). The isotopic composition of the primitive mantle, marked by a square, is normally estimated from this correlation and the Nd value of meteorites²⁶. These correlations are consistent with the relative compatibilities shown in Fig. 2, namely $\text{Rb} < \text{Sr}$, $\text{Sm} > \text{Nd}$, and $\text{Lu} > \text{Hf}$. Thus, mantle regions depleted in incompatible elements have low Rb/Sr and therefore low $^{87}\text{Sr}/^{86}\text{Sr}$ ratios, coupled with high Sm/Nd and Lu/Hf , and therefore high $^{143}\text{Nd}/^{144}\text{Nd}$ and $^{176}\text{Hf}/^{177}\text{Hf}$ ratios. Continental crust, which is enriched in incompatible elements, shows the opposite isotopic signatures. The position of OIB between the depleted MORB source and the continental crust suggests that the OIB sources might just be the result of back-mixing of various types of continental material into the mantle. It will be seen below that this is, for the most part, not the case.

Lead. The Pb isotope data (Fig. 4) are more complex than those for Sr, Nd and Hf, in part because the geochemical behaviour of lead is anomalous (see below and Fig. 2). The U/Pb ratio of the Earth cannot be inferred directly from meteorite data because the silicate portion of the Earth is severely depleted in lead (either because it has been lost from the Earth along with other volatile elements or because much of the lead has entered the core)²⁷. However, possible primitive mantle compositions are constrained to lie along a line (the 'geochron') determined by the age of the mantle. Also shown are estimates for the average upper (UCC) and lower (LCC) continental crust.

Two features are noteworthy. (1) The Pb data do not form a simple trend extending from depleted MORB through OIB to the continental crust. Thus, OIB sources cannot be explained simply by back-mixing of continental crust (as might have been inferred from the Sr–Nd–Hf data). (2) Given the relative compatibilities of U and Pb from Fig. 2, Pb isotope data in Fig. 4a for MORB should fall well to the left of the 'geochron' (locus of primitive-mantle Pb compositions). In other words, the Pb data are expected to show roughly the same topology as the Sr–Nd data in Fig. 3, but in fact most MORB data are roughly centred on the geochron, or even lie on its right-hand side. Moreover, the bulk continental crust (located somewhere between UCC and LCC) also lies close to the geochron rather than

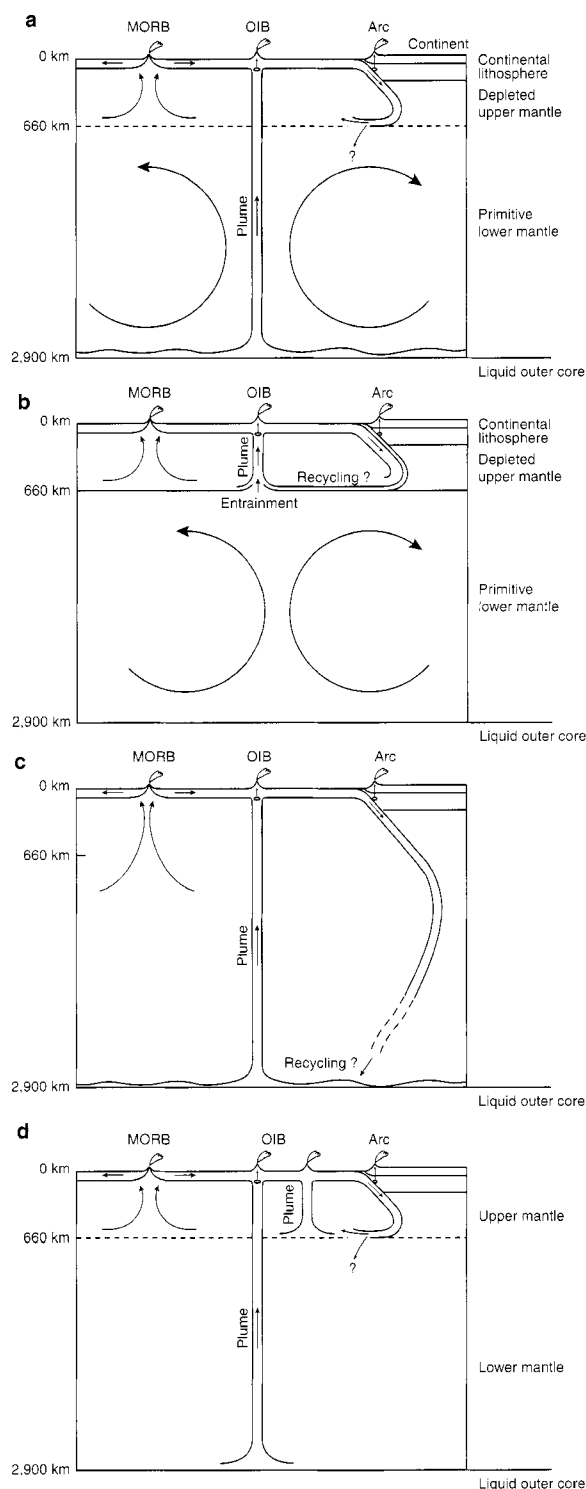


Figure 1 Models of mantle circulation. **a**, Old 'standard model' of two-layer circulation. The upper layer has been depleted in incompatible elements by the formation of continental crust. The layers are separated by either an endothermic phase change or by an intrinsic density contrast at a depth of 660 km. Plumes rise from the base of the lower, primitive (or less depleted) layer. The counterflow into the lower layer is not specified. (MORB, mid-ocean-ridge basalt; OIB, ocean island basalt.) **b**, Two-layer circulation with nearly complete isolation between upper and lower layers. Plumes rise from the base of the upper layer but may entrain small amounts of material from the lower layer. Plume sources are created by recycling of oceanic or continental lithosphere, which creates reservoirs that are enriched in incompatible elements. **c**, Whole-mantle, single-layer circulation with plumes rising from the core-mantle boundary. Plume sources are created as in **b**. **d**, Hybrid model, with circulation occurring primarily in two separate layers with small plumes rising from 660 km depth as in **b**, but with limited exchange by means of occasional foundering of subducted lithosphere and rise of strong plumes from the core-mantle boundary.

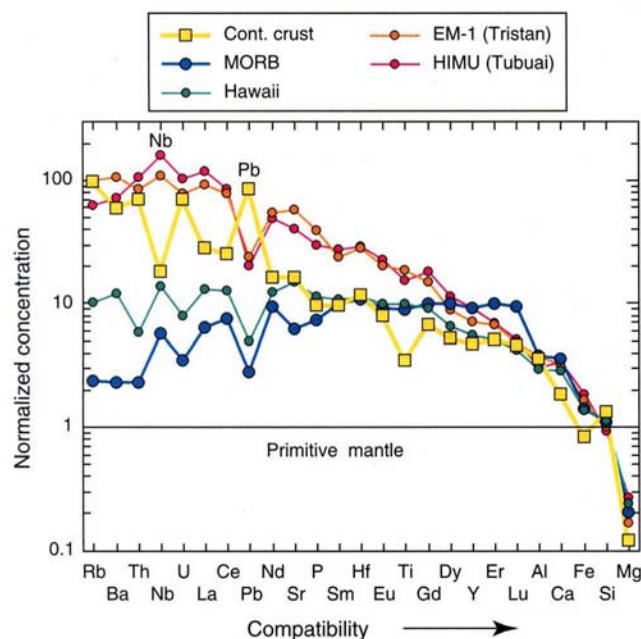


Figure 2 Concentrations of selected trace and major elements, arranged in the order of ascending compatibility and normalized to primitive-mantle concentrations, in average continental crust¹⁸, average MORB⁴, average Mauna Loa, Hawaii¹⁹, and three types of OIB: average Tristan and Inaccessible islands representing EM-1^{120,121}, and Tubuai representing HIMU islands⁴⁹. The patterns for MORB and the oceanic island basalts differ by their different enrichments (or depletions) in incompatible elements, but they are similar with respect to their positive Nb and negative Pb anomalies. The continental crust has opposite Nb and Pb anomalies. A fourth type of OIB, called EM-2, is similar to continental crust and has been omitted for clarity. (See text for explanations of EM-1, EM-2 and HIMU.)

far on its right-hand side. A solution to this puzzle (the 'lead paradox'²⁸) will be suggested below.

Correlations between Pb and other isotopes. Figure 5 shows relationships between $^{206}\text{Pb}/^{204}\text{Pb}$ and Sr isotope ratios. Although these isotope ratios correlate positively in MORB from the Atlantic and Pacific oceans, they do not correlate in MORB from the Indian Ocean and in the overall OIB–MORB array. This indicates that U–Pb and Th–Pb systems in the different mantle reservoirs evolve rather differently from the Rb–Sr (and Sm–Nd and Lu–Hf) systems.

In contrast, the parameter $^{208}\text{Pb}^*/^{206}\text{Pb}^*$ (the ratio of the radiogenic additions to the initial terrestrial lead, defined as $\{(^{208}\text{Pb}/^{204}\text{Pb}) - (^{208}\text{Pb}/^{204}\text{Pb})_{\text{init}}\} / \{(^{206}\text{Pb}/^{204}\text{Pb}) - (^{206}\text{Pb}/^{204}\text{Pb})_{\text{init}}\}$) does correlate with $^{143}\text{Nd}/^{144}\text{Nd}$ (and with Sr and Hf isotopes); see Fig. 6. This parameter depends primarily on Th/U (and age) but not on U/Pb or Th/Pb. The topology of Fig. 6 is rather similar to that of Fig. 3. This means that Th/U correlates with Sm/Nd (as well as with Lu/Hf and Rb/Sr) in mantle evolution, but U/Pb and Th/Pb do not. This identifies Pb (rather than Th or U) as the anomalous element²⁹. **Osmium.** Recent improvements in analytical methods have made osmium isotope analyses of oceanic basalts feasible. The primitive-mantle value of $^{187}\text{Os}/^{188}\text{Os}$ (or $^{187}\text{Os}/^{186}\text{Os}$) is identical or similar to the meteorite value^{30,31}. This means that the Earth's mantle has the same Re/Os ratio as meteorites, which is best explained by continued infall of meteoritic material after the main accretion and core formation of the Earth. Current results indicate the presence of higher-than-primitive $^{187}\text{Os}/^{188}\text{Os}$ in many OIB^{32–35}. Basalts have much higher (up to a factor of a hundred or more) Re/Os ratios than their mantle sources, so that growth of $^{187}\text{Os}/^{188}\text{Os}$ in the oceanic crust is quite rapid. Therefore, Os isotopes should furnish an

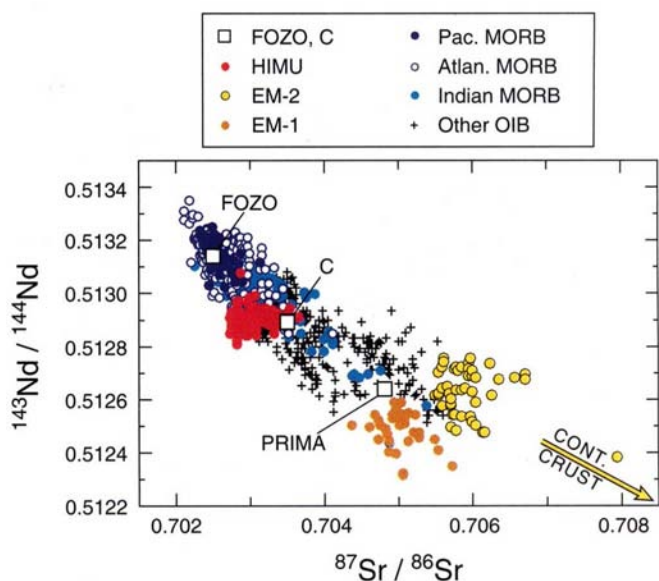


Figure 3 Nd and Sr isotopic compositions, assembled from literature sources, of MORB and OIB, with extreme HIMU, EM-1 and EM-2 samples marked in red, brown and yellow colours, respectively. Only those samples are shown for which Pb isotope data are also available (see Figs 4–6). HIMU samples are arbitrarily defined by having $^{206}\text{Pb}/^{204}\text{Pb} \leq 20$ (Fig. 4). EM-1 and EM-2 samples are also arbitrarily defined by their low $^{143}\text{Nd}/^{144}\text{Nd}$ and high $^{87}\text{Sr}/^{86}\text{Sr}$ values as shown on this diagram. The arrow points to the composition of average continental crust ($^{87}\text{Sr}/^{86}\text{Sr} = 0.72$, $^{143}\text{Nd}/^{144}\text{Nd} = 0.5118$). Also marked are the compositions representing the primitive mantle (PRIMA) and the proposed common mantle components of most plumes 'FOZO'¹⁶ and 'C'⁴⁰. See Supplementary Information for references used in compiling Figs 3–7.

excellent tracer for recycled basaltic material in the mantle and current results have generally been interpreted in this way. Mantle xenoliths from ancient subcontinental lithosphere have consistently lower-than-primitive $^{187}\text{Os}/^{188}\text{Os}$ values³⁶, and this constitutes an excellent potential tracer for recycled ancient lithosphere. However, no such low Os values have been found in oceanic basalts as yet.

Noble gas isotopes. The radiogenic nuclides ^4He and ^{40}Ar are produced by the decay of U and Th, and ^{40}K , respectively, and may be degassed from the Earth's interior into the atmosphere. The accumulation of ^{40}Ar in the atmosphere will be further discussed in the section on mass-balance considerations. Oceanic basalts also contain the non-radiogenic isotopes ^3He and ^{36}Ar (as well as Ne, Kr and Xe). Atmospheric helium is not recycled into the mantle because it is continually lost to space from the atmosphere, and nearly all ^3He now coming out the mantle is primordial, thus demonstrating that the Earth has never been completely degassed. (This nearly universally accepted view has been challenged by Anderson³⁷, who suggested that cosmic-dust-derived ^3He has been supplied to the mantle through subduction of pelagic sediments. However, high $^3\text{He}/^4\text{He}$ ratios do not correlate with other tracers of recycled subducted sediments in OIB³⁸, a fact that seriously weakens Anderson's argument.)

Relative to the atmospheric $^3\text{He}/^4\text{He}$ ratio ($R_A \equiv (^3\text{He}/^4\text{He})_{\text{atm}} = 1.4 \times 10^{-6}$), continental crust has low ratios ($^3\text{He}/^4\text{He} \approx 0.01 R_A$) because it is enriched in Th and U, which produce ^4He during decay. MORB have rather uniform values of $(8 \pm 1) R_A$ (ref. 39), and OIB range from 5 to $30 R_A$ (ref. 40). The existence of high- $^3\text{He}/^4\text{He}$ islands (Hawaii, Iceland, Bouvet, Galápagos, Easter, Juan Fernandez, Pitcairn, Samoa, Reunion and Heard islands; see, for example, ref. 40) is consistent with the layered-mantle model: plumes

might rise directly from the lower, high- $^3\text{He}/^4\text{He}$ reservoir through the upper, low- $^3\text{He}/^4\text{He}$ layer^{38,41–43}, or they might start from the base of the upper layer, in which case the helium must migrate into the plume source from the lower mantle^{2,44}.

Neon isotope data from basalts are still scarce because of experimental difficulties. Recent work has shown that Ne from the Hawaiian plume and from MORB have different $^{21}\text{Ne}/^{22}\text{Ne}$ and $^{20}\text{Ne}/^{22}\text{Ne}$ correlations, indicating the presence of a high- ^{20}Ne reservoir⁴⁵ in the plume source and a high- ^{21}Ne reservoir in the MORB source⁴⁶. Both components form mixing lines with a low- $^{21}\text{Ne}/^{22}\text{Ne}$ and $^{20}\text{Ne}/^{22}\text{Ne}$ reservoir of atmospheric origin. This provides additional evidence for a layered mantle with restricted exchange between layers.

Isotope taxonomy. The deviations from simple linear correlations in Figs 3–6 are not random but show systematic regularities^{1,14,47}, which are illustrated in these figures by colour coding. For example, all samples possessing $^{206}\text{Pb}/^{204}\text{Pb} \geq 20$ (Fig. 4) are shown as red dots and labelled 'HIMU' (see key in Fig. 3). They also form well-defined clusters on isotope diagrams not involving the element lead (Figs 3, 5, 6), even though they do not form an obvious endmember in them. Samples with this 'isotopic colour' occur in several oceanic islands in completely different parts of the world. This isotopic–chemical regularity demonstrates the existence of reproducible differentiation processes in the mantle and justifies the isotopic taxonomy discussed in the following paragraphs.

Zindler and Hart¹ suggested that the mantle contains the following isotopically extreme 'mantle components' (Figs 3–5), which contribute to the (usually mixed) sources of oceanic basalts. (1) HIMU ('high μ '; $\mu = ^{238}\text{U}/^{204}\text{Pb}$) has the highest Pb ratios and the lowest $^{87}\text{Sr}/^{86}\text{Sr}$ of any OIB (almost as low as MORB). Examples are St Helena, Austral Islands, Balleny Islands and the Azores. (2) EM-1 ('enriched mantle 1') occupies the lower left-hand corner of the Sr–Nd array (Fig. 3), and is concentrated in the right-hand corner of the $^{208}\text{Pb}^*/^{206}\text{Pb}^*$ –Nd array (Fig. 6), and the upper left-hand corner of the Pb–Pb array (Fig. 4b). Representatives are the Pitcairn and Tristan hotspots. (3) EM-2 ('enriched mantle 2') has the highest $^{87}\text{Sr}/^{86}\text{Sr}$ and relatively high $^{207}\text{Pb}/^{204}\text{Pb}$ ratios. Representatives are the Societies and Samoa hotspots. (4) DMM ('depleted MORB mantle') is not specifically marked on the figures. It is defined by the most depleted MORB samples with the highest $^{143}\text{Nd}/^{144}\text{Nd}$ and the lowest $^{87}\text{Sr}/^{86}\text{Sr}$, $^{206}\text{Pb}/^{204}\text{Pb}$, $^{207}\text{Pb}/^{204}\text{Pb}$ and $^{208}\text{Pb}/^{204}\text{Pb}$ ratios.

Many of the isotope data arrays for individual islands or hotspots are roughly linear, indicating that they represent mixtures of a more

enriched and a more depleted component originally thought to be the DMM. However, recently published data (for example, refs 38, 48, 49) show arrays in which DMM is clearly not a mixing end-member. This has led to the definition of a new component, called FOZO ('focal zone')¹⁶, whose composition is defined by the point of convergence, in three-dimensional isotope diagrams, of linear data arrays for individual ocean islands. A similar point of convergence of isotope arrays for MORB defines the component C ('common')⁴⁰. In the two-dimensional diagrams shown here, the convergence of HIMU and MORB arrays is most clearly seen in Fig. 4b.

There is much current debate about the meaning of this taxonomy. Authors disagree as to whether the linear isotopic trends seen in many hotspots are caused by the existence of two components in the plume source^{6,49}, whether isotopically distinct but internally uniform plume sources are mixed during ascent by entraining relatively depleted lower-mantle material^{16,18}, possibly producing a zoned plume⁵⁰, or whether similarly uniform plumes are mixed only with upper-mantle asthenosphere and/or lithosphere⁵¹. In general, individual hotspots contain only a limited range of isotopic compositions and most hotspots have distinct isotopic 'flavours'. On a much larger geographical scale, EM-1 and EM-2 compositions are concentrated in OIB just south of the Equator. This feature has been called the 'DUPAL anomaly'⁵² and has been correlated with large-scale seismic anomalies in the lowermost mantle. Also, nearly all Indian Ocean MORB have distinctly high $^{208}\text{Pb}/^{204}\text{Pb}$, $^{87}\text{Sr}/^{86}\text{Sr}$ and $^{208}\text{Pb}^*/^{206}\text{Pb}^*$ values relative to Pacific and Atlantic MORB (Figs 4–6). All of this provides strong evidence that the different isotopic flavours are not randomly distributed in the mantle but can be mapped. Nevertheless, conclusive interpretations are hard to come by because we cannot look into the mantle directly but must rely on simulations³³ or correlations between geophysical and geochemical observations. Until recently, the spatial resolution of both geophysical observations such as seismic tomography and numerical convection 'experiments' has been insufficient to 'see' or model relatively small features such as plumes. However, this resolution is improving rapidly and much progress can be expected within the coming decade.

Quite distinct from the spatial distribution of plume sources and other heterogeneities within the mantle is 'Darwin's question': what is the origin of the isotopic mantle species? To answer this, we must consider other aspects of mantle chemistry, in particular the concentration ratios of highly incompatible elements, which characterize the different species and should help to identify the

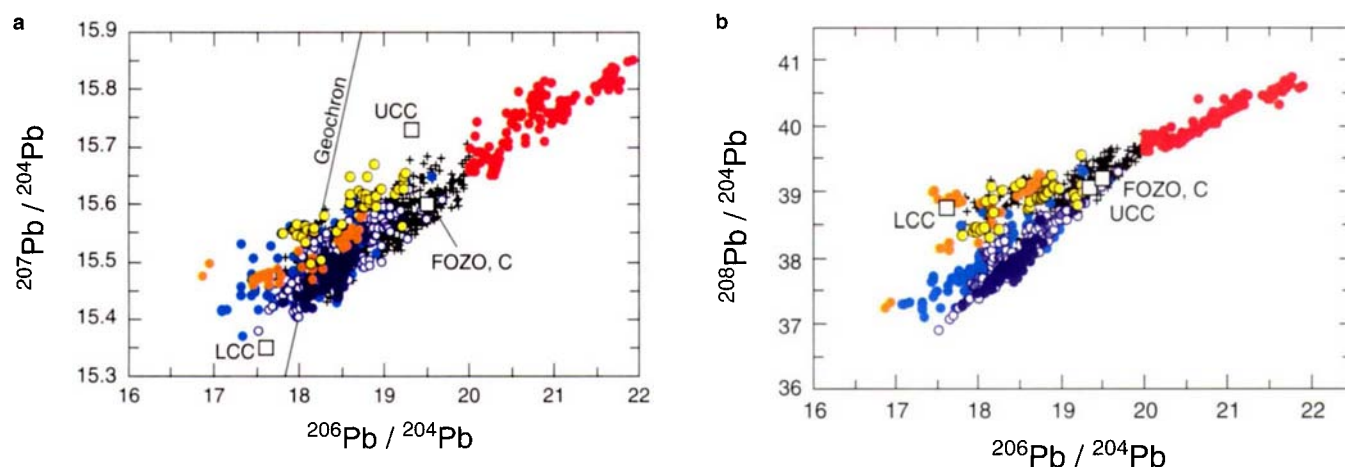


Figure 4 Pb isotopes for the samples shown in Fig. 3, using the same colour code. The solid line labelled "geochron" in **a** marks the locus of possible primitive mantle values assuming an overall age of the mantle of 4.50 Gyr (refs 27, 122). UCC

and LCC mark average compositions of upper and lower continental crust, respectively¹²³.

Table 1 Long-lived radioactive decay series used as tracers

Parent nuclide	Daughter nuclide	Half-life (yr)	Tracer ratio (radiogenic/ nonradiogenic)
^{147}Sm	^{143}Nd	106×10^9	$^{143}\text{Nd}/^{144}\text{Nd}$
^{87}Rb	^{87}Sr	48.8×10^9	$^{87}\text{Sr}/^{86}\text{Sr}$
^{176}Lu	^{176}Hf	35.7×10^9	$^{176}\text{Hf}/^{177}\text{Hf}$
^{187}Re	^{187}Os	45.6×10^9	$^{187}\text{Os}/^{188}\text{Os}$
^{40}K	^{40}Ar	1.25×10^9	$^{40}\text{Ar}/^{36}\text{Ar}$
^{232}Th	^{208}Pb	14.01×10^9	$^{208}\text{Pb}/^{204}\text{Pb}$
^{238}U	^{206}Pb	4.468×10^9	$^{206}\text{Pb}/^{204}\text{Pb}$
^{235}U	^{207}Pb	0.738×10^9	$^{207}\text{Pb}/^{204}\text{Pb}$

processes that ultimately created the isotopic differences.

Trace element ratios. Concentration ratios (hereafter called 'ratios') of incompatible trace elements can serve as tracers of source chemistry and complement the information gained from radiogenic isotopes. However, they must be used with care because, unlike isotope ratios, they may be fractionated during igneous processes. The ratio of two elements in a melt in equilibrium with its solid residue is given by

$$\frac{C_1}{C_2} = \frac{D_2(1-F) + F \cdot C_{1,0}}{D_1(1-F) + F \cdot C_{2,0}}$$

where C_1 and C_2 are the concentrations in the melt, $C_{1,0}$ and $C_{2,0}$ the concentrations in the original source, D_1 and D_2 the bulk partition coefficients, $D \equiv C_{\text{solid}}/C_{\text{liquid}}$, and F is the melt fraction. If $D_{1,2} \ll F$, that is, if the elements are highly incompatible or the melt fraction is large, the first term on the right-hand side vanishes and one obtains $C_1/C_2 \approx C_{1,0}/C_{2,0}$, so that the ratio in the melt will be the same as the ratio in the source. Such trace-element ratios can be just as useful as isotope ratios in serving as tracers of mantle chemistry.

Indicators of source depletion. The bulk partition coefficients of the most highly incompatible elements increase in the approximate order $\text{Ba} \approx \text{Rb} \leq \text{Th} < \text{U} \approx \text{K} < \text{La} < \text{Ce}$. Therefore, Ba/La, Rb/La, Th/U and Th/La ratios in basalts are generally greater than (or equal to) the respective source ratios. Thus, when these ratios are less than the primitive-mantle values, they document a previous depletion of the source rocks. This effect produces the leftward-descending branch in the MORB pattern seen in Fig. 2. A relative depletion for Rb and Ba is also seen in the Tubuai pattern; it is characteristic of all HIMU basalts and distinguishes them from EM-type basalts. It is consistent with the low Rb/Sr of HIMU sources inferred from their $^{87}\text{Sr}/^{86}\text{Sr}$ ratios, which indicate a 'depleted heritage' for both MORB and HIMU sources.

Globally 'uniform' ratios and the significance of Nb, Ta and Pb. Certain trace-element ratios are comparatively uniform in most oceanic basalts, irrespective of their absolute concentrations, which may range over two orders of magnitude from the most depleted MORB to the most enriched OIB. In such cases, the bulk partition coefficients of the two elements involved are nearly identical, or both are much smaller than the melt fraction F (see above)⁵⁴. Such uniform ratios include Ba/Rb, K/U (except in HIMU basalts), Nb/Ta, Zr/Hf, Y/Ho, Ti/Sm, Sn/Sm and P/Nd (see, for example, refs 55–57). These ratios are also similar in the average continental crust, and for the highly refractory element pairs Nb–Ta, Zr–Hf and Y–Ho, they are identical to the respective ('chondritic') ratios of the primitive mantle (see Box 1).

Some ratios, namely Nb/U ≈ 47 , Ta/U ≈ 2.7 and Ce/Pb ≈ 25 are similar in MORB and OIB, but their values are much higher than those of the continental crust and island arc volcanics (by factors of 4–5) and those estimated for the primitive mantle (by factors of 1.5–2.5)⁵⁴. The significance of this result is fourfold: (1) it shows that ocean island basalts do not sample predominantly primitive mantle reservoirs; (2) it links the origins of MORB and OIB sources; (3) it confirms independently the link between island arc magmatism and formation of continental crust as a consequence of plate

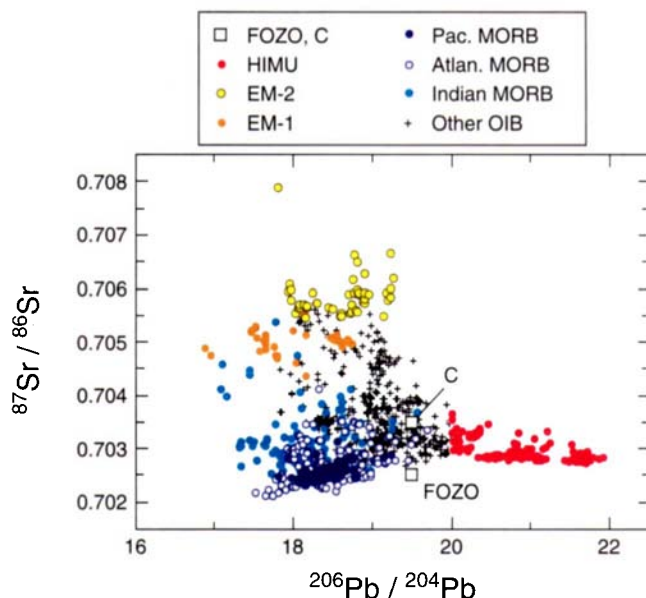


Figure 5 $^{87}\text{Sr}/^{86}\text{Sr}$ versus $^{206}\text{Pb}/^{204}\text{Pb}$ for the same samples as in Figs 3 and 4.

tectonic processes⁵⁸; (4) it provides excellent tracers to test for the presence of recycled continental crust in OIB sources.

Questions about the 'constancy' of these ratios and the above interpretation have recently been raised^{59,60}. The validity of the use of Nb/U as an important tracer of mantle sources is reaffirmed here by a new compilation, shown in Fig. 7, of Nb/U versus $^{143}\text{Nd}/^{144}\text{Nd}$ (expressed as ϵ_{Nd} ; see legend for definition). This shows average values for basalts from 33 ocean islands (or hotspots) and MORB, compiled from high-quality data published mostly since 1986. The great majority of all these basalt suites have average Nb/U between 40 and 60, independent of the source depletion or enrichment expressed by ϵ_{Nd} , and are in good agreement with the value of 47 ± 10 proposed in 1986⁵⁴. This is significantly higher than the primitive mantle ratio of Nb/U = 30 and the mean value of Nb/U ≈ 10 for continental crust and island arc rocks. Thus, most of the sampled mantle has been enriched in Nb (and Ta, which has similar properties to Nb) relative to U and other similarly incompatible elements. Possible mechanisms for this enrichment are discussed in the final section preceding the conclusion.

Figure 7 shows that normal OIB, including the HIMU islands and EM-1 islands, representing the full range of Nd-isotopic compositions in OIB, have Nb/U ratios very close to those of average MORB. Therefore, these OIB types cannot be generated by mixtures of depleted mantle and either primitive mantle or recycled continental crust (although the slightly low Nb/U ratios of EM-1 basalts permit the presence of a small sedimentary component; see below).

Obvious exceptions are the extreme EM-2 OIB from the Society and Marquesas hotspots (selected by having the highest $^{87}\text{Sr}/^{86}\text{Sr}$ at each hotspot) and the highly anomalous MORB from segment 3 of the Chile ridge (ref. 61 and E. Klein, unpublished data). This confirms that, except for EM-2, all types of oceanic basalt and their sources have very similar Nb/U values, which are significantly higher than primitive-mantle Nb/U. EM-2 OIB, and Chile ridge MORB are unique in that they, and only they, lie on mixing trajectories between DMM or HIMU mantle and average continental crust (probably in the form of mantle-recycled sediments).

An analogous but converse relationship holds for lead. Using Ce/Pb as a tracer, Hofmann *et al.*⁵⁴ and Newsom *et al.*⁶² showed that lead is systematically depleted in oceanic basalts, but overabundant in island arc rocks and continental crust (see also Fig. 2). A compilation of recent, more extensive data⁶³ shows that the bulk partition

coefficient of Pb is closer to that of Nd than to that of Ce. Therefore, although Ce/Pb (or better, Nd/Pb) ratios scatter considerably more than Nb/U ratios⁵⁹, they nevertheless appear to be good tracers of source ratios. For reasons of clarity, no example of an EM-2 basalt is shown on Fig. 2. However, on the Society Island (EM-2) hotspot, Ce/Pb and Nd/Pb ratios are anomalously low and well correlated with Nb/U and ⁸⁷Sr/⁸⁶Sr (refs 64–66), confirming the inference that EM-2 OIB, but not other OIB, contain small but significant amounts of recycled material of continental origin. The reasons for the anomalous behaviour of lead will be discussed in the section on the ‘lead paradox’ below.

Mass fraction of depleted mantle

By calculating how much of the mantle must be depleted in order to account for the incompatible-element enrichment in the continental crust, one can estimate how much of the mantle could still be chemically primitive. The mass fraction of the depleted mantle, X_d , is given by

$$X_d = \{X_c C_c (R_d - R_c) / (C_p (R_d - R_p))\} - X_c$$

where the subscripts d, c and p refer to depleted mantle, continental crust and primitive mantle, respectively, R is a ratio such as ¹⁴³Nd/¹⁴⁴Nd or Nb/U, and C is the concentration of the element in the denominator of R . Initial results based on Nd isotopes (for example, ref. 67) indicated depleted-reservoir sizes between 25% and 30% of the mantle. This corresponds to the mass of the mantle above the 660-km discontinuity and is thus consistent with the layered mantle model. However, other authors estimated the proportion of depleted mantle to be considerably larger, between 30% and 90% (for example, refs 1, 68–70), depending on specific assumptions about the composition of the continental crust and depending on the inclusion or omission of the OIB source reservoir(s).

Perhaps the best-constrained mass balance is that for radiogenic argon. Approximately 40% of the ⁴⁰Ar produced from ⁴⁰K in 4.5 Gyr of Earth history now resides in the atmosphere; only about 10% is in the continental crust and the depleted upper mantle, and the remaining 50% must still be in the lower mantle⁷¹. In terms of the above equation, this corresponds to $X_d = 0.5$. This is significantly greater than the mass fraction of the mantle above 660 km (about 0.27) but much less than the entire mantle. Thus, although the lower mantle is certainly no longer primitive, it is substantially less degassed, and thus probably also less depleted in incompatible elements, than the upper mantle (which has lost almost all of its ⁴⁰Ar). This constitutes a strong argument against simple whole-mantle convection, because such a circulation type would be expected to cause similar levels of degassing throughout the entire mantle.

Ages of mantle heterogeneities

The isotopic heterogeneities shown in Figs 3–6 require time to develop, particularly in view of the extremely long half-lives of the parent nuclides. If the parent/daughter ratio and the increase in radiogenic daughter nuclide were known, the age of a heterogeneity could be calculated. In practice, this is difficult for mantle domains, because the parent/daughter ratios of the mantle sources are generally not preserved in the basalts. Lead is an exception, because the production ratio of ²⁰⁷Pb/²⁰⁶Pb depends only on time, so that the slope of a straight line in ²⁰⁷Pb/²⁰⁴Pb–²⁰⁶Pb/²⁰⁴Pb space can be interpreted to yield the age of a single U–Pb differentiation event. Thus, if the OIB source regions had been formed in a single event, the slope of the correlation line in Fig. 4a would mean that this event occurred about 2 Gyr ago. However, continuous differentiation and remixing processes have been shown to produce data arrays with a similar slope (yielding an apparent age of 2 Gyr), even though the actual mean differentiation age of the individual mantle sources is only about 1–1.3 Gyr (refs 53, 72). Only the oldest members of the

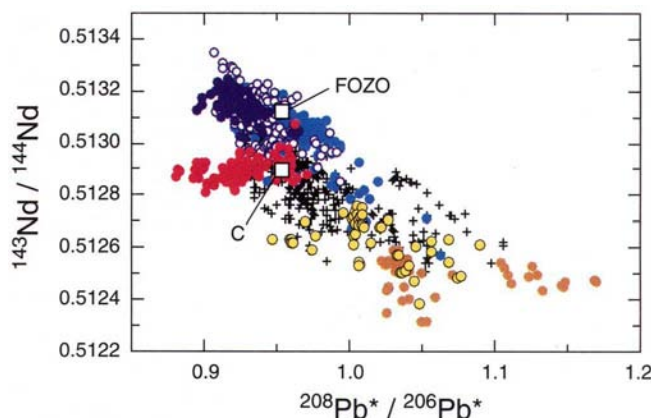


Figure 6 ¹⁴³Nd/¹⁴⁴Nd versus ²⁰⁸Pb*/²⁰⁶Pb* for the same samples as in Figs 3–5 (Here ²⁰⁸Pb*/²⁰⁶Pb* = {(²⁰⁸Pb/²⁰⁴Pb) – (²⁰⁸Pb/²⁰⁴Pb)_{init}} / {(²⁰⁶Pb/²⁰⁴Pb) – (²⁰⁶Pb/²⁰⁴Pb)_{init}}, where the subscript ‘init.’ designates the initial composition of the primitive mantle 4.5 Gyr ago).

club, the HIMU source regions with the highest ²⁰⁶Pb/²⁰⁴Pb values, were formed about 2 Gyr ago. Thus, the timescale observed for ‘oceanic’ mantle differentiation is likely to be significantly shorter than the mean age of the continental crust (about 2–2.5 Gyr; ref. 58).

Evaluation of geochemical mantle models

The large body of geochemical data reviewed above demonstrates that the Earth’s mantle has been partially depleted in incompatible elements by the formation of the continental crust, which has extracted roughly half of the total budget of highly incompatible elements from the mantle. This residual, depleted mantle is chemically heterogeneous on local, regional and global scales. Convective stirring has not homogenized the mantle, presumably because near-surface processes such as partial melting and subduction continue to reintroduce chemical and isotopic heterogeneities. The most highly depleted, and least heterogeneous region is located in the upper 660 km of the mantle and produces MORB by passive upwelling and partial melting. Plumes are more heterogeneous in composition and presumably originate from a boundary layer somewhere beneath this upper-mantle reservoir. There is a continuing debate about the existence of layered mantle convection. Layering may have been caused by an intrinsic, compositional density stratification that originated through crystallization of an early ‘magma ocean’ formed during segregation of the core⁷³. Current evidence from mineral physics⁷⁴ and experimental evidence on crystal–liquid partitioning of high-pressure minerals^{75,76} does not favour this model and probably precludes the survival of this type of mantle stratification. Layered convection may also be caused by the endothermic nature of the phase change from upper-mantle peridotite to a lower-mantle perovskite–wüstite mineralogy¹³. In this case, the major-element compositions of the upper and lower mantle may be the same, but the trace element and isotopic chemistry of the two reservoirs may differ drastically because of crust formation from the upper mantle (see section on mass fraction of depleted mantle).

Layered mantle, primitive OIB sources. This is the original ‘standard’ model (Fig. 1a) in its simplest form^{12,77}. Plumes rise from the primitive lower layer and mix with the upper, depleted layer on their way to the surface. This model is no longer tenable because (1) OIB sources require more than just two source components; (2) most OIB have distinctly non-primitive Pb isotopes to the right of the ‘geochron’ (Fig. 4a); (3) some OIB sources have lower-than-primitive ¹⁴³Nd/¹⁴⁴Nd ratios (Fig. 3); and (4) Nb/U and Ce/Pb (or Nd/Pb) ratios in OIB require non-primitive sources (Fig. 7).

Several authors have nevertheless recently proposed that some

OIB may be derived from primitive (or slightly depleted) sources (for example, refs 38, 42, 43). A particularly popular candidate is Hawaiian volcanism, especially the tholeiites from Oahu, Lanai and Kahoolawe. Basalts from these islands have close-to-primitive isotopic compositions of Sr, Nd, Hf and Pb (for example, ref. 78), and some Hawaiian volcanoes have high $^3\text{He}/^4\text{He}$ ratios^{41,79}. However, every Hawaiian volcano investigated so far shows the same Nb–Ta excess (and Pb deficiency)⁸⁰ that characterizes other OIB, and this precludes the derivation of Hawaiian volcanism from a primitive source.

In subsequent variants^{14,15,81–83} of the two-layer model, plumes originate from the base of the upper mantle, but they entrain small amounts of primitive (or slightly depleted) noble-gas-rich material from the lower mantle, which imparts the high $^3\text{He}/^4\text{He}$ and $^{20}\text{Ne}/^{22}\text{Ne}$ (Fig. 1b). There are two difficulties with these models. First it is not obvious why some plume types, particularly the HIMU types⁴⁰, should be consistently free of entrained, ^3He - and ^{20}Ne -rich lower-mantle material. Second, the amounts of lower-mantle material calculated from noble-gas evolution models are extremely low⁸³ and appear to be inconsistent with the mass of the less-depleted mantle derived from most recent mass-balance estimates (see above).

Other layered mantle models. Anderson (see, for example, refs 84, 85) has proposed some thought-provoking alternatives to the more conventional models. He suggested that the upper mantle is internally differentiated into a lower, garnetiferous, incompatible-element-depleted ‘piclogite’ layer above 660 km (the MORB source), and an upper (above 220 km), enriched peridotite layer, the ‘perisphere’, which is also the OIB source. Evidence for such an enriched layer is found in many young rifts, which typically deliver OIB-type magmas. However, the opening of the Red Sea shows that this enriched rift magmatism is soon replaced by normally depleted MORB magmatism, suggesting that the enriched perisphere is at best a thin veneer of enriched material that may well produce some of the unusual types of OIB, such as the Cameroon Line ‘hot line’ volcanoes⁵⁹, but could not sustain voluminous, long-lived, hotspot-type magmatism such as the Hawaiian–Emperor chain. Furthermore, there is scant evidence for the existence of such a perisphere in intra-oceanic regions, because laterally moving mid-ocean ridges, or ridge jumps, generally sample an ordinary, depleted MORB-type mantle rather than an enriched perisphere supposedly located in the uppermost mantle region. Finally, recent seismic imaging has traced the thermal anomaly beneath Iceland to a depth of 400 km and a diameter of about 100 km (ref. 86), indicating that the Iceland hotspot represents a mantle plume originating from at least 400 km depth. These features are difficult to reconcile with the perisphere model.

Whole-mantle convection models. In these models, the 660-km seismic discontinuity represents little or no barrier to convection. Geophysical support for this comes from evidence for deep, high-seismic-velocity structures (for example, ref. 87), interpreted as subducted lithosphere in the lower mantle, and from the fixed locations of hotspots, which appear to require that plumes originate in a lower-mantle region that has a much higher viscosity and moves much more slowly than does the upper mantle⁸⁸. It is also argued that flood basalt provinces, which are erupted during geologically very brief time spans and attain enormous volumes and lateral extent (about 2,000 km), are produced by plume heads that could have attained such dimensions only by thermal entrainment during plume ascent through the full depth of the mantle⁸⁹. Geochemical arguments are based on the so-called ‘FOZO’¹⁶ or ‘C’⁴⁰ isotopic component, which appears to be common to many plumes and differs isotopically from the MORB source. FOZO is thought to reside in the lower mantle where it is entrained by plumes rising from the core–mantle boundary¹⁶.

Metasomatic models. Metasomatism has been popular with many workers as a mechanism for enriching the upper mantle in incom-

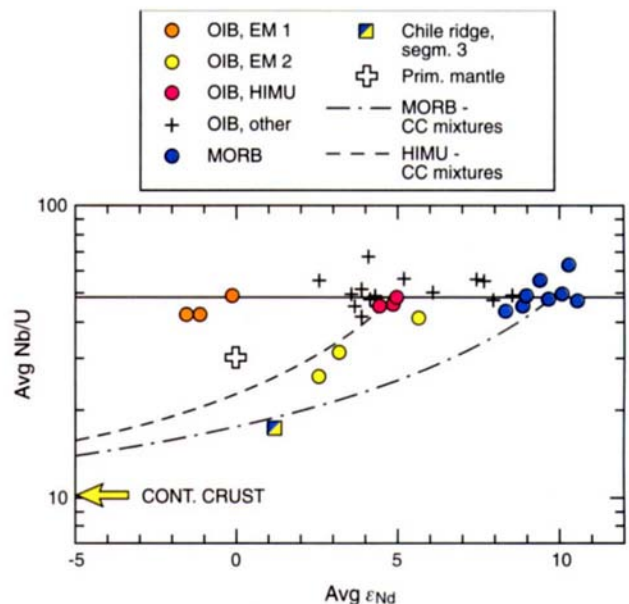


Figure 7 Nb/U plotted against ϵ_{Nd} , where ϵ_{Nd} is the deviation, in parts per 10⁴, of the $^{143}\text{Nd}/^{144}\text{Nd}$ ratio of a sample from that of present-day primitive mantle (p): $\epsilon_{\text{Nd}} = 10,000[(^{143}\text{Nd}/^{144}\text{Nd}) - (^{143}\text{Nd}/^{144}\text{Nd})_p] / (^{143}\text{Nd}/^{144}\text{Nd})_p$. The points shown are average values for suites of MORB and OIB using generally different samples but the same colour coding as in Figs 3–6. EM-1: Pitcairn, Tristan-Inaccessible, Kerguelen islands; EM-2: Society, Marquesas islands; HIMU: Tubuai, St Helena; other OIB: Cameroon Line islands, Iceland, Hawaii, Reunion, Comores, MacDonal sealamount, Easter hotspot, Bouvet, Gran Canaria, Tenerife, Azores, Cape Verde, Fernando di Noronha, Trinidad, Madeira, Ascension; MORB: Atlantic MORB 10°–23°N, Kolbeinsey ridge, Pacific MORB, Galapagos rift, Indian MORB, Chile ridge segment 3. All these oceanic basalt suites except EM-2 types have Nb/U averages similar to the global average of 47 ± 10 given by Hofmann *et al.*⁵⁴ using a much smaller set of data. Note that these Nb/U averages show no significant relationship to the degree of source enrichment or depletion as measured by Nd isotopes. Also, the primitive-mantle value lies completely outside the MORB–HIMU–EM-1 array. Only EM-2 type OIB and the segment 3 of the Chile ridge appear to be mixtures of MORB–HIMU type mantle with (recycled) material, probably sediments, from the continental crust. For references to this compilation, see Supplementary Information, except for unpublished data by C. H. Hémond *et al.* on Atlantic MORB 10°–23°N, and by E. Klein on the Chile ridge.

patible elements and volatiles, until it melts and generates OIB (see ref. 90). Vollmer⁹¹ has based a global mantle model on such a mechanism. The effects of metasomatism are certainly evident in many mantle xenoliths⁹², but the scale of these processes is unknown and may be small relative to the source volume of OIB⁸⁰. The viability of the plume model has largely obviated the ‘need’ for metasomatic mechanisms to generate OIB directly.

Metasomatism may also enrich substantial volumes of the mantle above a subduction zone. The oceanic crust is hydrated by ridge-crest hydrothermal and sea-floor alteration processes, and it is covered by wet sediments. During subduction, this hydrous material is heated and dehydrated. The escaping fluids migrate through the overlying mantle, thereby metasomatizing it. Island arc volcanism is likely to remove much of the effects of this metasomatism^{93,94} (as incompatible elements are partitioned into the fluid and ultimately into the melt), but some of the metasomatized mantle may be incorporated in the subcontinental lithosphere or returned to the convecting part of the mantle. In either case, the metasomatized material may ultimately be recycled through the mantle and reappear in intra-oceanic volcanism (see below).

A variant of this mechanism is metasomatism by the migration of small amounts of melt from the asthenosphere into the oceanic or

continental lithosphere^{3,59}, followed by recycling as above⁹⁵. All of these metasomatic models are difficult to test by observations, because the only tangible evidence for them is found in fist-size or smaller mantle xenoliths, which yield no clues as to the quantities of material transported or the distances travelled.

Recycling models. The idea that recycling crust through the mantle could explain the isotopic and chemical diversity of oceanic basalts goes back almost 30 years⁹⁶ but has found wide acceptance only in the past few years, because mantle convection was previously thought to be highly efficient at rehomogenizing the chemical and isotopic heterogeneities introduced into the mantle during subduction.

Continental crust. Armstrong^{96,97} proposed that the continental crust has maintained a constant volume through a steady-state process of growth (through accretion of island arcs) and destruction through subduction of continent-derived sediments, and suggested that this could also explain the isotopic diversity of oceanic basalts. The geochemical evidence from Pb isotopes and from Nb/U and Ce/Pb ratios now rules out large amounts of recycled continental material in OIB sources (see above), except in the cases of EM-2 OIB and Indian Ocean MORB, which have low Nb/U and Ce/Pb values correlated with high ⁸⁷Sr/⁸⁶Sr and ²⁰⁷Pb/²⁰⁴Pb ratios, and which can be explained by additions of small amounts (1–2%) of recycled sediments to the source mantle^{6,49,66,98} (see Fig. 7).

Oceanic crust. Subduction of oceanic lithosphere continuously recycles oceanic crust and thus injects about 20 km³ of enriched material per year into the mantle. This is the uniformitarian basis of the recycling model of refs 99 and 100. Soon after subduction, this crust is transformed into eclogite (and later to higher-pressure equivalents), which is about 2–4% denser than ordinary mantle (depending on pressure⁵³). The assumed segregation of the dense former oceanic crust from its lithospheric base and the main mantle mass is inhibited at normal mantle viscosities, but is enhanced in the hot boundary layer at the base of the convection system (660 or 2,900 km). Storage and isotopic evolution of some of this material for periods of 1–2 Gyr, followed ultimately by incorporation into plumes rising from this boundary layer, could generate the MORB–HIMU isotope array⁵³.

Strongly altered oceanic crust is evidently not being recycled, because alteration adds many 'mobile' elements such as Rb and U (but not Th), which would cause isotopic changes quite unlike the observed OIB isotopic arrays¹⁰¹. If such alteration has been ubiquitous in the past, it must therefore be assumed that the altered, hydrous parts of the crust were efficiently stripped from the system during subduction, so that the 'additive' effects of hydrothermal alteration were removed. Chemical and isotopic evidence from subduction-related volcanism supports this^{93,94}. In contrast, the hydrothermal subtraction of certain elements, such as lead, from the oceanic crust is probably irreversible, and this should contribute to the high U/Pb and Th/Pb ratios responsible for the extremely radiogenic lead of HIMU basalts^{49,102}.

The model of oceanic crustal recycling is supported by the still sparse isotope data for osmium. HIMU basalts in particular have elevated ¹⁸⁷Os/¹⁸⁸Os ratios, which can be explained by an ancient basaltic source component^{32–35}. Other oceanic basalt types, such as Hawaiian basalts, which have Pb isotopic compositions closer to MORB, also have high ¹⁸⁷Os/¹⁸⁸Os ratios⁵⁰ and may represent recycled, deeper oceanic crust depleted in Th and U (ref. 103).

Subcontinental lithosphere. Old subcontinental lithosphere is comparatively cold and dense, so it may be delaminated and sink into the mantle to become a plume source⁸¹. Because it has been 'pre-aged' isotopically, while it was still attached to the crust, it does not need an extended period of storage in the convecting part of the mantle before it acquires the isotopic signatures found in plumes. This is a popular source candidate for EM-1 type OIB because some xenoliths derived from the subcontinental mantle have isotopic signatures similar to that of EM-1. The lithosphere may well be

enriched as a result of fluid metasomatism during subduction^{104,105} or melt migration from the asthenosphere⁹⁵, but the scale of such processes remains largely a matter of speculation. This recycling hypothesis can now be tested using osmium isotopes, because ancient lithosphere is unique in having ¹⁸⁷Os/¹⁸⁸Os significantly lower than primitive-mantle values³⁶. So far, Os data for only two EM-1 type basalts have been published, and both of these have high, rather than low, ¹⁸⁷Os/¹⁸⁸Os (ref. 33).

Continental basalts are also commonly thought to sample the subcontinental lithosphere (see, for example, refs 104, 106) and they often resemble EM-1 type OIB. However, recent realization that continental flood basalts may sample the heads of deep-mantle plumes makes the argument somewhat circular. Thus it is not clear whether the EM-1 character is due to contamination of a plume head by the continental lithosphere or whether this signature was part of the original plume source¹⁰⁷.

Towards a unified model for mantle evolution

My preferred model for the evolution of the mantle is based on a simple-minded, uniformitarian approach that uses known geological processes and avoids exotic processes whenever possible. Thus, I assume that much of the chemical diversity of the isotopic mantle 'species' is caused by subduction, which introduces oceanic crust and small amounts of sediments of variable composition into the mantle.

Timescales of crust–mantle evolution. Continental crust formation extracts incompatible elements very efficiently from the mantle and enriches them in the crust roughly 100-fold. Continental crust may be returned to the mantle by subduction of sediments, but its lifetime is at least 20 times that of the oceanic crust, which has a mean life of only about 100 Myr. The present-day production rate of continental crust (1 km³ yr^{−1}; ref. 108) is also about 20 times lower than that of oceanic crust (20 km³ yr^{−1}). Because of the grossly different timescales, global differentiation may be approximated roughly by a two-stage process⁵⁴:

(1) The 'continental stage' or 'primary differentiation' of primitive mantle into continental crust with high Rb/Sr and Th/U, and low Sm/Nd, Nb/U and Ce/Pb ratios, and a residual mantle with complementary low Rb/Sr and Th/U, and high Sm/Nd, Nb/U and Ce/Pb. The mean age of the crust thus produced is 2–2.5 Gyr (ref. 58). The trace-element patterns of MORB (Fig. 2) reflect the residual nature of their source mantle. (2) The 'oceanic stage' continuously differentiates the residual mantle by ridge and subduction processes and rehomogenizes it by convection. This differentiation–homogenization process has time constants averaging 1–1.3 Gyr^{53,72} (see discussion on ages of mantle heterogeneities), and it creates most of the chemical and isotopic diversity seen in Figs 3–6.

The complementary Nb(Ta) and Pb anomalies in the oceanic and continental crusts (Fig. 2) are caused by special processes during continent production. Niobium and tantalum are retained in the mantle by Ti minerals¹⁰⁹ or amphibole¹¹⁰ during subduction-related volatile transfer¹¹¹ and/or magmatism. The continental crust grows, in part, by accretion of subduction-related, Nb–Ta depleted island-arc magmas and thus acquires the large negative Nb–Ta and smaller Ti anomalies seen in Fig. 2. The mantle therefore becomes relatively enriched in Nb and Ta, and this positive anomaly is homogenized through relatively rapid (see above) convective stirring. This explains the relatively uniform Nb/U ratios seen in Fig. 7. Oceanic (MORB and OIB) magmatism does not create new Nb–Ta anomalies because the minerals or volatiles that could fractionate Nb/U are absent in the MORB–OIB melting region.

'Lead paradox'. The positive lead anomaly in the continental crust and the complementary negative anomalies in oceanic basalts (Fig. 2) are created by hydrothermal processes in two stages. First, lead is extracted from the basaltic crust and redeposited in metalliferous sediments by hydrothermal ridge processes¹⁰². When the crust is subducted, additional lead is extracted from the crust and

the sediments by dehydration. This lead is ultimately transferred to the continental crust via island arc magmatism (see above).

This transfer process provides the key to the solution of the 'lead isotope paradox' and the absence of a global correlation of $^{206}\text{Pb}/^{204}\text{Pb}$ with Sr and Nd isotopes (Fig. 5). The hydrothermal processes have enriched the continental crust in lead to a similar degree as in uranium (see Fig. 2). Consequently, the continental crust has, on average, similar U/Pb and therefore $^{206}\text{Pb}/^{204}\text{Pb}$ ratios as the (residual) mantle. Similarly, most of the MORB data also lie close to the 'geochron' in Fig. 4a, rather than falling to the 'depleted', low- $^{206}\text{Pb}/^{204}\text{Pb}$ side (caused by low U/Pb) that would be expected if the sources were produced by purely magmatic depletion processes. (For a detailed discussion of the 'best' position of the geochron, see ref. 27.) Continent-derived ocean-floor sediments have $^{206}\text{Pb}/^{204}\text{Pb}$ ratios similar to average upper oceanic crust (Fig. 4), but possess an extremely large range of U/Pb ratios¹¹², which will, in time, generate a large range of Pb-isotopic compositions. When such sediments are recycled through the mantle, they will destroy the correlations between the U/Pb, Rb/Sr and Sm/Nd ratios that are produced by purely magmatic differentiation processes. This appears to have happened in EM-2 (and possibly also in EM-1) type OIB, as well as in Indian Ocean MORB, which show no correlations between Sr (or Nd) and Pb isotopes (Fig. 5). In contrast, Pacific and Atlantic MORB show good correlations and appear to be nearly free of recycled continental material.

Role of plumes and mantle layering. Geochemical evidence, especially that derived from noble gases, favours some type of layered mantle convection, but one in which the layers are not completely isolated. I therefore suggest that plumes may come from both the 660-km discontinuity and the core-mantle boundary. Small, short-lived plumes and those carrying no enrichment in primordial ^3He are most likely to have a shallow origin. Large, long-lived, ^3He -rich plumes, and especially those that have generated enormous volumes of oceanic or continental flood basalts, are more likely to have come from the base of the lower mantle⁸⁹.

This history of deep-plume activity may well have been episodic and linked to times of high rates of continent formation^{18,108}. Several of the large oceanic plateaux were created in the Cretaceous period, and this may constitute an unusual event in Earth history¹¹³. These plateaux may represent the heads of deep-mantle plumes⁸⁹ that will ultimately be accreted to the continental crust. The relatively large number of ^3He -rich plumes may therefore be mostly remnants of deep-mantle plumes that started in Cretaceous time and are still active today. Similarly, deep subduction of oceanic lithosphere into the lower mantle, for which there is mounting seismological evidence⁸⁷, may be a sporadic event that balances the occasional deep-mantle plume flux¹⁷.

In geologically 'normal' times, both plate-tectonic convection and plume generation may be confined to the upper mantle. This implies further that at present we are seeing both the remnants of a deep-plume and deep-subduction event (in the form of the remaining plume trails) and the resumption of isolated upper-mantle convection dominated by plate tectonics and upper-mantle plumes¹⁸.

Depletion, replenishment, recycling. The uppermost mantle, the source of MORB, is depleted in incompatible elements by two processes. (1) Devolatilization and melting during subduction transfer incompatible elements to long-time (or permanent) storage in the continental crust and lithosphere; and (2) a portion of the subducted oceanic crust is stored at the base of the convecting system and thus sequesters incompatible elements temporarily. The remainder of this crust is mixed back into the MORB reservoir. Consequently, the MORB mantle has a 'marble cake' structure²⁴ containing schlieren of enriched, recycled material. In addition, plume injection creates 'regional' (100–1,000 km) scale chemical and isotopic gradients in MORB mantle (see, for example, ref. 12). The Indian Ocean MORB source region differs from Pacific and

Atlantic regions in that it has been 'polluted' by ancient sediments⁹⁸ or subcontinental lithosphere¹¹⁴ (see section on the lead paradox).

The former oceanic crust stored as potential plume sources at 660 and 2,900 km depth contains a wide variety of 'enriched' material. HIMU plume sources are enriched in former MORB, in which hydrothermal ridge processes had caused loss of lead and therefore an increase in (U, Th)/Pb (refs 6, 49). EM-2 sources represent recycled oceanic crust containing a few per cent of continent-derived sediment^{49,66,115}. This is consistent with their high $^{207}\text{Pb}/^{204}\text{Pb}$ and $^{87}\text{Sr}/^{86}\text{Sr}$ and low Nb/U and Ce/Pb ratios (see Fig. 7). EM-1 sources are the most enigmatic members of the mantle zoo. Their slightly subnormal Nb/U, low $^{206}\text{Pb}/^{204}\text{Pb}$ and high $^{208}\text{Pb}/^{206}\text{Pb}$ ratios are consistent with a contribution from either recycled, ancient pelagic sediment^{64,9} or recycled ancient subcontinental lithosphere^{81,107,116}. Osmium isotopes (see above) should resolve the issue.

Conclusion and outlook

Although many parts of the story remain controversial and difficult to decipher, I regard the following conclusions to be reasonably robust. Recycling of oceanic crust and lithosphere plays an important role in generating mantle heterogeneities. Available geochemical mass balances favour compositional layering, which is maintained by partial or intermittent convective isolation between upper and lower mantle.

The controversy over the issue of convective layering and depth of plume sources is likely to be resolved when the so far rather fuzzy seismic imaging of the Earth's interior becomes comparable in resolution to that achieved by geochemical mapping, so that geophysical and geochemical data can be more specifically correlated. The origin of the geochemical 'mantle species' will be clarified by a rapidly expanding data base of osmium isotopes and high-quality data for well-established tracer elements such as Nb and Pb, as well as relatively 'new' tracers such as B and Sb, and by high-precision oxygen isotope microanalysis of igneous minerals¹¹⁷, capable of recognizing a sedimentary or hydrothermal prehistory of the magma. □

A. W. Hofmann is at the Max-Planck-Institut für Chemie, Postfach 3060, 55020 Mainz, Germany

1. Zindler, A. & Hart, S. *Annu. Rev. Earth Planet. Sci.* **14**, 493–571 (1986).
2. Allègre, C. J. *Earth Planet. Sci. Lett.* **86**, 175–203 (1987).
3. Sun, S.-S. & McDonough, W. F. in *Magmaism in the Ocean Basins* (eds Saunders, A. D. & Norry, M. J.) 313–345 (Spec. Publ., Geological Soc., Oxford, 1989).
4. Hofmann, A. W. *Earth Planet. Sci. Lett.* **90**, 297–314 (1988).
5. Olson, P., Silver, P. G. & Carlson, R. W. *Nature* **344**, 209–215 (1990).
6. Weaver, B. L. *Earth Planet. Sci. Lett.* **104**, 381–397 (1991).
7. Davies, G. F. & Richards, M. A. *J. Geol.* **100**, 151–206 (1992).
8. Carlson, R. W. *Rev. Geophys.* **32**, 337–361 (1994).
9. White, W. M. *Rev. Geophys. (suppl.)*: US Natl Rep. to Int. Union Geodesy and Geophysics 1991–1994, 19–24 (1995).
10. Wilson, J. T. *Can. J. Phys.* **41**, 863–870 (1963).
11. Morgan, W. J. *Nature* **230**, 42–43 (1971).
12. Schilling, J.-G. *Nature* **242**, 565–571 (1973).
13. Christensen, U. R. *Annu. Rev. Earth Planet. Sci.* **23**, 65–87 (1995).
14. Allègre, C. J. & Turcotte, D. L. *Geophys. Res. Lett.* **12**, 207–210 (1985).
15. Galer, S. J. G. & O'Nions, R. K. *Nature* **316**, 778–782 (1985).
16. Hart, S. R., Hauri, E. H., Oschmann, L. A. & Whitehead, J. A. *Science* **256**, 517–520 (1992).
17. Machetel, P. & Weber, P. *Nature* **350**, 55–57 (1991).
18. Stein, M. & Hofmann, A. W. *Nature* **372**, 63–68 (1994).
19. Whitehead, J. A. & Luther, D. S. *J. Geophys. Res.* **80**, 705–717 (1975).
20. Loper, D. E. *Tectonophysics* **187**, 373–384 (1991).
21. Batiza, R. *Earth Planet. Sci. Lett.* **60**, 195–206 (1982).
22. Halliday, A. N., Dickinson, A. P., Fallick, A. E. & Fitton, J. G. *J. Petrol.* **29**, 181–211 (1988).
23. Griffiths, R. W. & Campbell, I. H. *Earth Planet. Sci. Lett.* **99**, 66–78 (1990).
24. Allègre, C. J. & Turcotte, D. L. *Nature* **323**, 123–127 (1986).
25. Hofmann, A. W. & Hart, S. R. *Earth Planet. Sci. Lett.* **39**, 44–62 (1978).
26. Jacobsen, S. B. & Wasserburg, G. J. *Earth Planet. Sci. Lett.* **67**, 137–150 (1984).
27. Galer, S. J. G. & Goldstein, S. L. in *Earth Processes: Reading the Isotopic Code* (eds Basu, A. & Hart, S. R.) 75–98 (Am. Geophys. Union, Washington DC, 1996).
28. Allègre, C. J. *Earth Planet. Sci. Lett.* **5**, 261–269 (1969).
29. Allègre, C. J., Dupre, B. & Lewin, E. *Chem. Geol.* **56**, 219–277 (1986).
30. Allègre, C. J. & Luck, J. M. *Earth Planet. Sci. Lett.* **48**, 148–154 (1980).
31. Meisel, T., Walker, R. J. & Morgan, J. W. *Nature* **383**, 517–520 (1996).
32. Hauri, E. H. & Hart, S. R. *Earth Planet. Sci. Lett.* **114**, 353–371 (1993).
33. Reisberg, L. et al. *Earth Planet. Sci. Lett.* **120**, 149–167 (1993).
34. Roy-Barman, M. & Allègre, C. J. *Earth Planet. Sci. Lett.* **129**, 145–161 (1995).
35. Marcantonio, F., Zindler, A., Elliott, T. & Staudigel, H. *Earth Planet. Sci. Lett.* **133**, 397–410 (1995).

36. Pearson, D. G., Carlson, R. W., Shirey, S. B., Boyd, F. R. & Nixon, P. H. *Earth Planet. Sci. Lett.* **134**, 341–357 (1995).
37. Anderson, D. L. *Science* **261**, 170–176 (1993).
38. Farley, K. A., Natland, J. H. & Craig, H. *Earth Planet. Sci. Lett.* **111**, 183–199 (1992).
39. Allegre, C. J., Moreira, M. & Staudacher, T. *Geophys. Res. Lett.* **22**, 2325–2328 (1995).
40. Hagan, B. & Graham, D. *Science* **272**, 991–995 (1996).
41. Kurz, M. D., Jenkins, W. J., Hart, S. R. & Clague, D. *Earth Planet. Sci. Lett.* **66**, 388–406 (1983).
42. Kellogg, L. H. & Wasserburg, G. J. *Earth Planet. Sci. Lett.* **99**, 276–289 (1990).
43. Porcelli, D. R. & Wasserburg, G. J. *Geochim. Cosmochim. Acta* **59**, 4921–4937 (1995).
44. O'Nions, R. K. & Tolstikhin, I. N. *Earth Planet. Sci. Lett.* **124**, 131–138 (1994).
45. Honda, M., McDougall, I., Patterson, D. B., Douleris, A. & Clague, D. A. *Geochim. Cosmochim. Acta* **57**, 859–874 (1993).
46. Sarda, P., Staudacher, T. & Allegre, C. J. *Earth Planet. Sci. Lett.* **91**, 73–88 (1988).
47. White, W. M. *Geology* **13**, 115–118 (1985).
48. Barling, J. & Goldstein, S. L. *Nature* **348**, 59–62 (1990).
49. Chauvel, C., Hofmann, A. W. & Vidal, P. *Earth Planet. Sci. Lett.* **110**, 99–119 (1992).
50. Hauri, E. H., Lassiter, J. C. & DePaolo, D. J. *J. Geophys. Res.* **101**, 11793–11806 (1996).
51. Class, C., Goldstein, S. L., Galer, S. J. G. & Weiss, D. *Nature* **362**, 715–721 (1993).
52. Hart, S. R. *Nature* **309**, 753–757 (1984).
53. Christensen, U. R. & Hofmann, A. W. *J. Geophys. Res.* **99**, 19867–19884 (1994).
54. Hofmann, A. W., Jochum, K.-P., Seufert, M. & White, W. M. *Earth Planet. Sci. Lett.* **79**, 33–45 (1986).
55. Hofmann, A. W. & White, W. M. *Z. Naturforsch.* **38**, 256–266 (1983).
56. Jochum, K.-P., Hofmann, A. W., Ito, E., Seufert, H. M. & White, W. M. *Nature* **306**, 431–436 (1983).
57. Jochum, K. P., Hofmann, A. W. & Seufert, H. M. *Geochim. Cosmochim. Acta* **57**, 3585–3595 (1993).
58. Taylor, S. R. & McLennan, S. M. *The Continental Crust: its Composition and Evolution*, 1–312 (Blackwell Scientific, Oxford, 1985).
59. Halliday, A. N. *et al.* *Earth Planet. Sci. Lett.* **133**, 379–395 (1995).
60. Sims, K. W. W. & DePaolo, D. J. *Eos* **72**, 526–527 (1991).
61. Klein, E. M. & Karsten, J. L. *Nature* **374**, 52–57 (1995).
62. Newsom, H. E., White, W. M., Jochum, K. P. & Hofmann, A. W. *Earth Planet. Sci. Lett.* **80**, 299–313 (1986).
63. Hofmann, A. W. *Abstracts of the 49th IAVCEI Meeting* (Geol. Soc. Australia, Canberra, 1993).
64. Devey, C. W. *et al.* *J. Geophys. Res.* **95**, 5049–5066 (1990).
65. Hémond, C., Devey, C. W. & Chauvel, C. *Chem. Geol.* **115**, 7–45 (1994).
66. White, W. M. & Duncan, R. A. in *Earth Processes: Reading the Isotopic Code* (eds Basu, A. & Hart, S. R.) 183–206 (Am. Geophys. Union, Washington DC, 1996).
67. Jacobsen, S. B. & Wasserburg, G. J. *J. Geophys. Res.* **84**, 7411–7427 (1979).
68. Davies, G. F. *Nature* **290**, 208–213 (1981).
69. Allegre, C. J., Hart, S. R. & Minster, J.-F. *Earth Planet. Sci. Lett.* **66**, 191–213 (1983).
70. Hofmann, A. W. *Phil. Trans. R. Soc. Lond. A* **328**, 425–439 (1989).
71. Allegre, C. J., Hofmann, A. W. & O'Nions, K. *Geophys. Res. Lett.* (in the press).
72. Allegre, C. J. & Lewin, E. *Earth Planet. Sci. Lett.* **136**, 629–646 (1995).
73. Anderson, D. L. *Science* **223**, 347–355 (1984).
74. Jackson, I. & Rigden, S. M. in *The Earth's Mantle: Composition, Structure and Evolution* (ed. Jackson, I.) (Cambridge Univ. Press, Cambridge, in the press).
75. Kato, T., Ringwood, A. E. & Irifune, T. *Earth Planet. Sci. Lett.* **89**, 123–145 (1988).
76. Ohtani, E., Kawabe, I., Moriyama, J. & Nagata, Y. *Contrib. Mineral. Petrol.* **103**, 263–269 (1989).
77. Sun, S. S. & Hanson, G. N. *Geology* **3**, 297–302 (1975).
78. Roden, M. F., Trull, T., Hart, S. R. & Frey, F. A. *Geochim. Cosmochim. Acta* **58**, 1431–1440 (1994).
79. Kurz, M. D. & Kammer, D. P. *Earth Planet. Sci. Lett.* **103**, 257–269 (1991).
80. Hofmann, A. W. *Chem. Geol.* **57**, 17–30 (1986).
81. McKenzie, D. & O'Nions, R. K. *Nature* **301**, 229–231 (1983).
82. Galer, S. J. G., Goldstein, S. L. & O'Nions, R. K. *Chem. Geol.* **75**, 252–290 (1989).
83. O'Nions, R. K. & Tolstikhin, I. N. *Earth Planet. Sci. Lett.* **139**, 213–222 (1996).
84. Anderson, D. L. *Earth Planet. Sci. Lett.* **57**, 13–24 (1982).
85. Anderson, D. L. *Rev. Geophys.* **33**, 125–149 (1995).
86. Wolfe, C. J., Bjarnason, I. T., VanDecar, J. C. & Solomon, S. C. *Nature* **385**, 245–247 (1997).
87. Grand, S. P. *J. Geophys. Res.* **99**, 11591–11621 (1994).
88. Steinberger, B. & O'Connell, R. J. *Geophys. J. Int.* (submitted).
89. Campbell, I. H. & Griffiths, R. W. *Earth Planet. Sci. Lett.* **99**, 79–93 (1990).
90. Bailey, D. K. *Nature* **296**, 525–530 (1982).
91. Vollmer, R. *Geology* **13**, 452–454 (1983).
92. Menzies, M. A. & Hawkesworth, C. J. *Mantle Metasomatism*, 465 (Academic, London, 1987).
93. Ryan, J. G. & Langmuir, C. H. *Geochim. Cosmochim. Acta* **57**, 1489–1498 (1993).
94. Ishikawa, T. & Nakamura, F. *Nature* **370**, 205–208 (1994).
95. McKenzie, D. & O'Nions, R. K. *J. Petrol.* **36**, 133–159 (1995).
96. Armstrong, R. L. *Rev. Geophys.* **6**, 175–199 (1968).
97. Armstrong, R. L. *Phil. Trans. R. Soc. Lond. A* **301**, 443–472 (1981).
98. Rehämper, M. & Hofmann, A. W. *Earth Planet. Sci. Lett.* (submitted).
99. Hofmann, A. W. & White, W. M. *Yb Carnegie Inst. Wash* **79**, 477–483 (1980).
100. Hofmann, A. W. & White, W. M. *Earth Planet. Sci. Lett.* **57**, 421–436 (1982).
101. Hart, S. R. & Staudigel, H. in *Crust/Mantle Recycling at Convergence Zones* (eds Hart, S. R. & Gülen, L.) 15–28 (Kluwer, Dordrecht, 1989).
102. Peucker-Ehrenbrink, B., Hofmann, A. W. & Hart, S. R. *Earth Planet. Sci. Lett.* **125**, 129–142 (1994).
103. Hofmann, A. W. & Jochum, K. P. *J. Geophys. Res.* **101**, 11831–11839 (1996).
104. Carlson, R. W. *Geochim. Cosmochim. Acta* **48**, 2357–2372 (1984).
105. Hawkesworth, C. J., Rogers, N. W., van Calsteren, P. W. C. & Menzies, M. A. *Nature* **311**, 331–335 (1984).
106. Carlson, R. W. *Austr. J. Earth Sci.* **38**, 525–544 (1991).
107. Mahoney, J., Nicollet, C. & Dupuy, C. *Earth Planet. Sci. Lett.* **104**, 350–363 (1991).
108. Reymer, A. & Schubert, G. *Tectonics* **3**, 63–77 (1984).
109. Green, T. H. *Chem. Geol.* **117**, 1–38 (1994).
110. Ionov, D. A. & Hofmann, A. W. *Earth Planet. Sci. Lett.* **131**, 341–356 (1995).
111. Keppeler, H. *Nature* **380**, 237–240 (1996).
112. Ben Othman, D., White, W. M. & Patchett, J. *Earth Planet. Sci. Lett.* **94**, 1–21 (1989).
113. Larson, R. L. *Geology* **19**, 547–550 (1991).
114. Mahoney, J. J. *et al.* *J. Geophys. Res.* **94**, 4033–4052 (1989).
115. White, W. M. & Hofmann, A. W. *Nature* **296**, 821–825 (1982).
116. Milner, S. C. & le Roex, A. P. *Earth Planet. Sci. Lett.* **141**, 277–291 (1996).
117. Eiler, J. M. *et al.* *Earth Planet. Sci. Lett.* (in the press).
118. Rudnick, R. L. & Fountain, D. M. *Rev. Geophys.* **33**, 267–309 (1995).
119. Hofmann, A. W. *et al.* *Eos* **74**, 630 (1993).
120. Le Roex, A. P., Cliff, R. A. & Adair, B. J. I. *J. Petrol.* **31**, 779–812 (1990).
121. Cliff, R. A., Baker, P. E. & Mateer, N. J. *Chem. Geol.* **92**, 251–260 (1991).
122. Faure, G. *Principles of Isotope Geology* 1–589 (Wiley, New York, 1986).
123. Zartman, R. E. & Haines, S. M. *Geochim. Cosmochim. Acta* **52**, 1327–1339 (1988).
124. Jagoutz, E. *et al.* *Proc. 10th Lunar Planet. Sci. Conf., Geochim. Cosmochim. Acta (Suppl. 11)* **2**, 2031–2050 (1979).
125. McDonough, W. F. & Sun, S.-S. *Chem. Geol.* **120**, 223–253 (1995).

Supplementary Information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from Mary Sheehan at the London editorial office of Nature.

Acknowledgements: I acknowledge unpublished Nb–U data made available by F. Albarède (Reunion), C. Hémond (Atlantic MORB) and E. Klein (Chile ridge). I thank R. Carlson and W. White for formal reviews; C. Allegre, R. Carlson, R. Cliff, S. Galer, S. Goldstein, S. R. Hart, K. P. Jochum and W. McDonough for informal reviews and/or discussions of the issues involved; and H. Hartshorn for reading the manuscript and suggesting numerous improvements in style and clarity. This article was completed while the author was a visiting investigator at the IPGP in Paris as recipient of the MESR-A. v. Humboldt Award.

Correspondence should be addressed to A.W.H. (e-mail: hofmann@geobar.mpg.de).

Like-charge attractions in metastable colloidal crystallites

Amy E. Larsen* & David G. Grier

The James Franck Institute and Department of Physics, The University of Chicago, 5640 S. Ellis Avenue, Chicago, Illinois 60637, USA

Sub-micrometre charged latex spheres can be suspended in water to form regular arrays known as colloidal crystals. In contrast to most conventional solids, colloidal crystals can be forced into metastable superheated states. The structure and dynamics of these metastable crystals show evidence for strong, long-range attractions between the similarly charged spheres. Such attractive interactions are inconsistent with the accepted theory of colloidal interactions, and might influence the properties of many natural and industrial suspensions.

Charge-stabilized colloidal suspensions of extremely uniformly sized polymer spheres were developed in the 1950s as media for paints and other surface coatings. These suspensions have attracted considerable attention in recent years for their utility as model systems with which to study the mechanisms of structural phase transitions¹. Depending on their concentration and chemical environment, colloidal spheres can form up into regular crystalline arrays or devolve into fluid disorder. Transformations between the ordered and disordered states are phase transitions analogous to melting and freezing of atomic matter. Unlike atoms, however, colloidal spheres can be tracked with a conventional light microscope. Their structural transformations thus provide unparalleled opportunities to investigate the microscopic mechanisms of phase transitions.

Here we describe how the study of phase transitions sheds new light on the fundamental properties of colloidal suspensions. In particular, we find evidence in an exotic colloidal melting transition for long-range attractive interactions between like-charged microspheres. When combined with direct measurements of the spheres' pairwise interaction potential, these observations strongly suggest that such attractive interactions may be responsible for a variety of other unexplained phenomena observed over the past decade in bulk colloidal suspensions.

The most striking of these anomalies include large stable voids in otherwise homogeneous suspensions^{2,3} and equilibrium phase separation between colloidal fluids of different densities^{4,5}. Neither should be possible in a system with purely repulsive interactions, although explanations based on impurity effects⁶ have proved difficult to exclude. Persistent quantitative discrepancies between predicted^{7–11} and observed^{12,13} melting points of colloidal crystals are no less disturbing, but might reflect relatively minor shortcomings in the theory. The long-ranged attractions we observe would account naturally for these effects, but are qualitatively inconsistent with the long-accepted theory for colloidal interactions. We suggest, therefore, that the theory for colloidal interactions requires substantial revision.

Colloidal interactions and phase transitions

The experiments described below were carried out on suspensions of polystyrene sulphate spheres of diameter $2a = 0.652 \pm 0.005 \mu\text{m}$ dispersed in deionized water (catalogue number 5065A, Duke Scientific, Palo Alto, CA). These spheres are synthesized with a large number of ionizable sulphate salt groups chemically bonded to their surfaces. These groups dissociate in water, each leaving a single negative charge bound to its sphere's surface and a compensating positively charged counterion in solution. One sulphate group subtends roughly 10 nm^2 so that a micrometre-scale sphere has a titratable charge

of the order of 10^5 electron equivalents. Not all surface groups dissociate at once, though, so the effective charge number is typically much smaller. While these specially prepared spheres have consistent and highly uniform surface charges, many naturally occurring and industrially prepared colloids also acquire static surface charges¹. Results discussed here should pertain to them as well.

The interaction between an isolated pair of similarly charged spheres is purely repulsive and has been shown both theoretically^{14,15} and experimentally^{16–18} to adopt the screened-Coulomb form

$$\frac{U(r)}{k_B T} = Z^2 \lambda_B \left[\frac{\exp(\kappa a)}{1 + \kappa a} \right]^2 \frac{\exp(-\kappa r)}{r} \quad (1)$$

where r is the spheres' centre-to-centre separation. This expression for the interaction energy was originally derived by Derjaguin, Landau, Verwey and Overbeek (DLVO)^{14,15} and accounts approximately for the counterions' exclusion from the spheres' interiors. The range of the interactions is limited by the concentration n of simple ions in the water and is described by the Debye–Hückel screening length $\kappa^{-1} = (4\pi n \lambda_B)^{-\frac{1}{2}}$, where $\lambda_B = e^2/\epsilon k_B T$ is the characteristic separation between ions carrying a single electronic charge, e , at temperature T in a fluid of dielectric constant ϵ .

The complete DLVO theory includes a term accounting for van der Waals attractions. This term is neglected in the following discussion because it contributes less than $0.01 k_B T$ to the interaction potential for polystyrene microspheres immersed in water and separated by more than 100 nm (ref. 19).

The suspension at an initial volume fraction around $\phi = 0.02$ is contained in the space between two parallel glass walls separated by $28 \mu\text{m}$. The edges of this sample volume are hermetically sealed

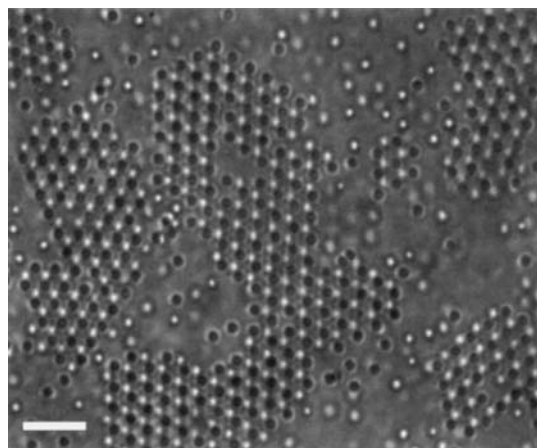


Figure 1 Metastable colloidal crystallites adrift in a colloidal fluid, after 10 minutes of melting. Scale bar, $5 \mu\text{m}$; field of view, $42 \times 35 \mu\text{m}$.

* Present address: Department of Physics, St Olaf College, 1520 St Olaf Avenue, Northfield, Minnesota 55057, USA.

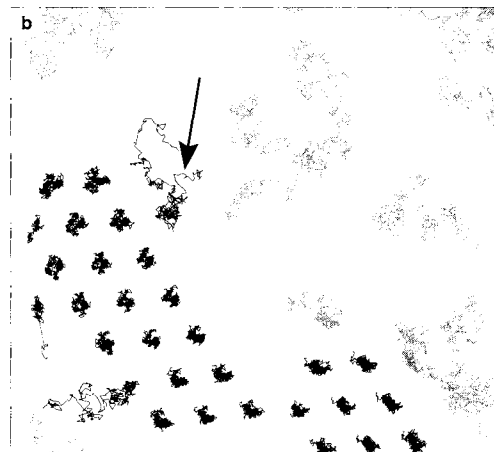
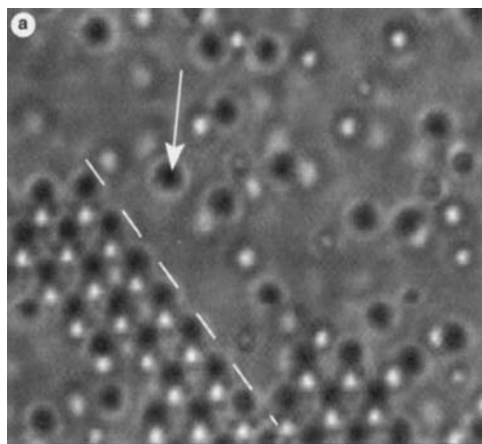


Figure 2 a, Video micrograph of the surface plane of a faceted colloidal crystal in contact with a dilute colloidal fluid. The dashed line demarks a (110) facet of the crystallite, while the arrow points out a sphere about to attach to the surface.

b, Trajectories of the spheres in **a** taken over 15 s. Heavy lines show trajectories that are part of the crystal, light lines show trajectories in the fluid.

against contamination by airborne CO_2 and make diffusive contact with reservoirs of ion exchange resin. We estimate the steady-state concentration of residual stray ions in this system to be less than $n = 5 \times 10^{-6}$ M. Such low ionic strengths provide for long-ranged electrostatic interactions with Debye–Hückel screening lengths extending to $\kappa^{-1} = 400$ nm.

If we assume that the interactions among a collection of spheres can be built up by superposing pairwise repulsions of the form described by equation (1), then the phase diagram for the ensemble of spheres is well understood^{7–11}. In particular, a fluid interacting through screened-Coulomb repulsions can be frozen to a crystal of face-centred cubic (f.c.c.) or body-centred cubic (b.c.c.) symmetry by increasing the volume fraction of spheres in the suspension while leaving other parameters fixed. Transformations qualitatively consistent with this picture have been observed experimentally^{12,13} in charge-stabilized colloidal suspensions.

Conventional crystals are rigid enough to support themselves against gravity and so determine their own density. Colloidal crystals exist only within containers. Were they not confined, we might expect mutually repelling spheres to drift apart until the forces causing them to order were overwhelmed by randomizing thermal agitation. At this point, the crystal would melt. We have found, however, that some colloidal crystals made of like-charged spheres maintain their integrity even without confining walls. The structure and dynamics of these metastable crystals provide strong evidence for the influence of long-ranged attractions not accounted for in the DLVO theory.

Superheated crystals

We take advantage of forces which arise during low-frequency electrophoresis to compress a colloidal fluid past its freezing point²⁰. The latent heat of freezing is absorbed by the surrounding water whose heat capacity is comparatively so enormous that the suspension's temperature does not change. When the electric field is turned off, the suspension is free to return to its equilibrium fluid state so that any remaining crystals are superheated with respect to the fluid.

Parallel-plate gold electrodes patterned onto one of the sample volume's glass surfaces provide electrical contact to the suspension. A 60 Hz signal, 10 V peak-to-peak, is applied across the 1-mm gap and drives the spheres by electrophoresis. The interplay of this force and electro-osmotically driven bulk fluid flow engenders a Magnus force which pushes the spheres towards the walls and away from the sample volume's mid-plane. Similar forces drive spheres settling under gravity to the walls of fluidized beds²¹. The resulting increase in density induces freezing at the walls and produces several polycrystalline layers.

As we have previously described²⁰, superheated colloidal crystals in suspensions of moderate ionic strength melt at a rate limited only by diffusion of spheres away from the crystal surface. All traces of order in such suspensions typically disappear within 10 seconds. More thoroughly deionized suspensions at lower ionic strength display dramatically different behaviour, retaining residual metastable crystallites drifting through the fluid for as long as an hour. The examples in Fig. 1 are in a non-equilibrium ordered state 10 minutes after compression was stopped. The largest of these crystallites consist of five layers of spheres in a face-centred cubic packing. Individual spheres appear either light or dark depending on their distance from the focal plane of the microscope's $100 \times \text{NA } 1.4$ oil immersion objective. Dark spheres in Figs 1 and 2a constitute the layer closest to the glass wall, the brightest are in focus in the second layer, and the third layer is barely visible as light smudges. Nearest-neighbours are separated by $d = 1.8 \mu\text{m}$ or roughly three diameters.

That the melting rate changes dramatically with changing ionic concentration is not necessarily surprising. Increasing the screening length could push the ensemble of spheres close to their crystal–fluid coexistence line. In that case, the free-energy difference driving spheres from the crystal back into the fluid could be vanishingly small and the melting process correspondingly slow. Analysis of the crystals' structure, however, shows that other forces are at work.

Stable facets

The most prominent features of crystals such as those in Figs 1 and 2a are their well defined facets. A crystal's facets are stable against thermal capillary waves only if the latent heat of freezing per particle exceeds

$$L > 4k_{\text{B}}T \frac{z}{\eta} \quad (2)$$

where η is the coordination number of a particle at the surface and z is the bulk coordination number²². For low-index facets such as that in Fig. 2a, $z/\eta \approx 2$, so that the latent heat is at least $L = 8k_{\text{B}}T$ per particle.

This lower limit already is a factor of eight larger than the largest latent heat anticipated for spheres with purely repulsive pairwise interactions. Entropically stabilized hard-sphere crystals with short-range contact repulsions are predicted²³ to have a latent heat only as large as $1k_{\text{B}}T$ per particle. Longer-range repulsions lead to even smaller latent heats²⁴. The discrepancy with our observations is accounted for most easily by positing the existence of a long-range attractive interaction contributing to the crystal's cohesive energy in addition to the core electrostatic repulsion. The attraction cannot be attributed to van der Waals interaction as such fluctuation-induced forces are utterly negligible at the crystal's lattice spacing¹⁹. To account for the facets' stability, each nearest-neighbour 'bond' would have to reduce the system's free energy by of the order of $0.5k_{\text{B}}T$.

Interfacial kinetics

The kinetics of attachment and detachment at the interface provide further evidence of attractive pairwise interactions. Figure 2b shows the computer-measured trajectories¹⁷ followed by the spheres photographed in Fig. 2a. Spheres in the crystal simply rattle about their lattice positions with transit times comparable to the 1/30-second interval between video frames. For the most part, spheres in the neighbouring fluid roam freely.

The track we have emphasized near the centre of Fig. 2b follows a sphere which begins its trajectory as part of the fluid but joins the crystal interface for $\tau = 6$ s before resuming its random walk. In this time, the sphere wanders no further than $r_0 = 0.25 \mu\text{m}$ from its lattice position. The probability $P(r_0|\tau) = 1 - \exp(-r_0^2/(4D\tau))$ that a brownian sphere with the fluid's measured¹⁷ diffusion coefficient of $D = 0.15 \mu\text{m}^2 \text{s}^{-1}$ would remain so well localized is less than 0.02. Rather than being localized by chance, the sphere appears to be held in place by an energetic barrier to desorption over which it eventually escapes. If we estimate of the order of $N = 100$ unsuccessful attempts to leave the surface by thermal excitation, then the barrier height is at least $E_b = k_B T \ln N \approx 4k_B T$. the existence of a surface activation energy also would explain why only one other sphere detaches from the crystal during the 15 seconds for which the trajectories were traced.

Crystals made from mutually repulsive spheres should put up no barrier to detachment. The intersphere attraction of around $0.5k_B T$ per neighbour deduced from the crystals' faceted structure, however, would account naturally for the estimated E_b .

Smallest metastable clusters

As the initially large metastable crystallites melt, they create a population of smaller crystals. Eventually, these crystals shrink to the radius R^* at which their surface energy exceeds the binding energy holding them together and they dissociate. The critical radius for a face-centred cubic crystal with lattice constant d , latent heat per particle L , and surface tension γ is

$$R^* = \sqrt{2} \frac{\gamma d^3}{L} \quad (3)$$

The smallest metastable crystals thus allow us to set a lower limit on the surface tension of the crystal–fluid interface.

Figure 3 shows four stages in the dissociation of a minute crystal whose radius is one lattice constant. Very shortly after one additional sphere separates from the surface, the remaining spheres go on their separate ways. Although we regularly observe crystals such as that in Fig. 3a, we do not find smaller long-lived ordered clusters. If we identify $R^* = d$ and apply the above result for L , then we can estimate a lower limit for the surface tension: $\gamma > L/(\sqrt{2}d^3) \approx 5.6k_B T d^{-2}$. This value is about an order of magnitude larger than the largest surface tension anticipated for a system with purely repulsive pairwise interactions^{25–27}. Once again, the discrepancy is naturally explained by assuming a long-range attractive component in the pairwise interaction.

Measuring the interaction

These three experimental observations suggest that the metastable colloidal crystallites are bound together by a strong long-range attraction not accounted for by the DLVO theory. Recent direct measurements of the pairwise colloidal interaction potential^{16–18} demonstrate that no such attraction appears for isolated pairs of spheres. Attractions are observed, however, when spheres are rigidly confined to the plane by closely spaced parallel glass surfaces^{18,28,29}, although the mechanism for this attraction is not yet understood. As the metastable crystals form at glass walls, a wall-mediated interaction might explain our observations.

Two considerations stand in the way of this interpretation. First, the wall separation in the present experiments is roughly ten times larger than in the experiments for which pair attractions were

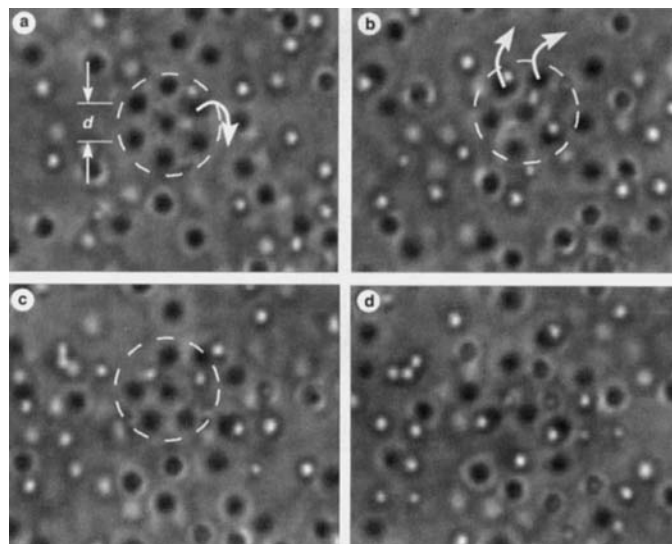


Figure 3 Metastable cluster of ten spheres against a glass wall. The appearance of clusters such as this, and the lack of smaller clusters, suggest that the maximum unstable radius is $R^* \approx 1d$. **a**, The initial ten-sphere cluster indicated by a dashed circle. The arrow indicates the direction in which a sphere will eventually leave the surface. **b**, The same cluster 6 seconds later when one sphere detaches. The crystallite has drifted in the interval. The next two spheres to detach are indicated by arrows. **c**, After another 2 seconds, the cluster is reduced to seven spheres. **d**, 1 second later, the crystal has completely dissipated.

measured, and the ability of a single wall to induce attractive interactions has not been demonstrated. Second, the walls' measured influence on colloidal interactions falls off with distance, h , from the wall¹⁸. Consequently, the effect might not extend far enough to stabilize the crystalline layers farthest from the wall.

To examine the range and magnitude of intersphere attractions mediated by a single glass wall, we performed direct interaction measurements in the same sample cell used to study the metastable crystals. The measurement technique is described in detail in ref. 16. Two optical tweezers are used to position a pair of spheres at fixed separations in the focal plane of the microscope. When the laser beam powering the traps is interrupted, the spheres move under the combined influence of random thermal forces and their mutual interaction. Blinking the traps on and off in synchrony with the video camera's shutter allows us to sample experimentally the markovian propagator for the master equation whose steady-state solution is the equilibrium pair distribution function for the spheres, $g(r)$. The pair distribution function is related to the pair interaction potential through Boltzmann's equation:

$$g(r) = \exp\left(-\frac{U(r)}{k_B T}\right) \quad (4)$$

Approximately 20,000 images of sphere pairs taken at 1/30-s intervals suffice to measure an interaction potential with 60-nm spatial resolution and $0.2k_B T$ energy sensitivity over a range of $6 \mu\text{m}$ and $5k_B T$, respectively. To avoid many-body contributions to the measurements, we diluted the suspension so that no other spheres wandered within $25 \mu\text{m}$ of the test pair.

The lower trace in Fig. 4 was obtained with the optical tweezers focused $h = 2.5 \pm 0.5 \mu\text{m}$ from the sample cell's upper glass wall. At this distance from the wall, the measured interaction potential has a strongly attractive minimum at a centre-to-centre separation of $3.5 \mu\text{m}$. This confirms that a single wall can induce an attraction between like-charged colloidal spheres comparable to that deduced from structural analysis of the metastable crystals. The range and strength of the attraction are too great to be accounted for by van der Waals attraction. Image charges in the nearby glass walls

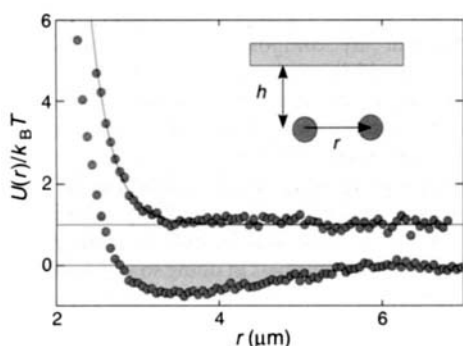


Figure 4 Interaction potential as a function of centre-to-centre separation for a pair of 0.65- μm -diameter polystyrene sulphate spheres near a charged glass surface. The experimental geometry is represented schematically in the inset. The upper curve was measured a distance $h = 9.5 \pm 1.0 \mu\text{m}$ from the nearest wall and has been offset by $1k_B T$ vertically for clarity. The dashed line is a fit to equation (1). The lower curve was acquired at $h = 2.5 \pm 0.5 \mu\text{m}$ and has an attractive minimum $\sim 0.7k_B T$ deep for spheres separated by more than five diameters.

similarly would exert too weak an influence. The attraction's dependence on distance from the wall suggests instead a counterion-mediated mechanism¹⁸.

No attraction is evident in the upper trace of Fig. 4 which was measured $h = 9.5 \pm 1.0 \mu\text{m}$ from the lower wall. As has been observed for spheres confined by a pair of closely spaced glass walls¹⁸, attractive interactions only appear when the spheres' ion clouds interact strongly with the glass wall's screening charge. These conditions are not considered in the formulation of the DLVO theory^{14,15}. The DLVO theory should apply, however, for isolated pairs of spheres far from the walls. The dashed curve in Fig. 4 is a three-parameter fit to equation (1) for the effective charge $Z = 7,300$, the Debye-Hückel screening length $\kappa^{-1} = 275 \text{ nm}$, and an additive offset.

The absence of an attractive minimum at $h = 9.5 \mu\text{m}$ suggests that the wall-induced attractive interaction alone cannot stabilize the fourth and fifth layers of crystals such as those in Fig. 1 which are at least $h = 7 \mu\text{m}$ from the wall. Furthermore, diluting the suspension for the interaction measurements probably reduced the electrolyte's ionic strength and increased not only the screening length but also the range of the wall-induced interaction. This seems likely because the centre-to-centre separation of both the attractive minimum and the repulsive core in $U(r)$ exceed the crystals' nearest-neighbour spacing of $1.8 \mu\text{m}$. In the higher ionic concentrations at which the metastable crystals were studied, the influence of the wall is likely to extend even less far than is indicated in Fig. 4. Interaction measurements in the confined geometry by Kepler and Fraden²⁸ also show that the range and strength of the wall-induced interaction depend strongly on the electrolyte's ionic concentration. They confirm, in particular, that higher ionic strengths lead to shorter-ranged attractions.

Attractive phenomenology

When combined with the results of direct interaction measurements, our observations on metastable colloidal crystals lead us to several insights into colloidal interactions. The structure and dynamics of the metastable crystals provide clear evidence for strong, long-range attractions acting between like-charged colloidal microspheres which are not explained by the standard DLVO theory. The attractive pair potentials measured in the presence of confining glass walls are strong enough to account for the crystals' properties, but do not extend far enough from the walls to stabilize their full three-dimensional structure. Their stability, therefore, suggests that the planes of charged spheres in the crystal act in a manner comparable to charged glass planes in providing the geometrical confinement apparently necessary for attractive interactions. A charged wall with its counterions indeed may resemble a

concentration of charge-stabilized spheres in a coarse-grained or mean-field description. If this is the case, then confinement-induced attractive interactions might arise in charge-stabilized colloidal suspensions even without glass walls. The nature and mechanism of this and the wall-mediated attraction are open questions and pose a formidable challenge to the theory of colloidal interactions.

Discrepancies between predictions of the DLVO theory and measurements on charge-stabilized colloidal suspensions have fuelled an active debate about alternatives to the DLVO theory for colloidal interactions. Such alternatives include the Sogami-Ise theory^{30,31}, which describes long-range pairwise attractions mediated by counterion distributions at least qualitatively similar to our observations. Very recent direct interaction measurements¹⁸, however, show that the Sogami-Ise theory does not correctly describe the interactions between unconfined pairs of spheres. It probably does not, therefore, explain the present observations.

The existence and behaviour of superheated colloidal crystals suggest instead that many-body effects may contribute an attractive component to the pairwise interactions between charged microspheres. Given the many useful, interesting and economically important properties which microscopic interactions impart to colloidal suspensions, this result has significance beyond the immediate application to non-equilibrium phase transitions. $\square$

Received 24 July; accepted 18 November 1996.

- Russel, W. B., Saville, D. A. & Schowalter, W. R. *Colloidal Dispersions* (Cambridge Univ. Press, 1989).
- Ito, K., Yoshida, H. & Ise, N. Void structure in colloidal crystals, *Science* **263**, 66–68 (1994).
- Doshio, S. *et al.* Recent study of polymer latex dispersions, *Langmuir* **9**, 394–411 (1993).
- Tata, B. V. R., Rajalakshmi, M. & Arora, A. K. Vapour-liquid condensation in charged colloidal suspensions, *Phys. Rev. Lett.* **69**, 3778–3781 (1992).
- Tata, B. V. R. & Arora, A. K. Tata and Arora reply, *Phys. Rev. Lett.* **72**, 787 (1994).
- Palberg, T. & Würth, M. Comment on 'Vapour-liquid condensation in charged colloidal suspensions', *Phys. Rev. Lett.* **72**, 786 (1994).
- Kremer, K., Robbins, M. O. & Grest, G. S. Phase diagram of Yukawa systems: model for charge-stabilized colloids, *Phys. Rev. Lett.* **57**, 3694–2697 (1986).
- Shih, W. Y., Aksay, I. A. & Kikuchi, R. Phase diagrams of charged colloidal particles, *J. Chem. Phys.* **86**, 5127–5132 (1987).
- Robbins, M. O., Kremer, K. & Grest, G. S. Phase diagram and dynamics of Yukawa systems, *J. Chem. Phys.* **88**, 3286–3312 (1988).
- Meijer, E. J. & Frenkel, D. Melting line of Yukawa system by computer simulation, *J. Chem. Phys.* **94**, 2269–2271 (1991).
- Stevens, M. J. & Robbins, M. O. Melting of Yukawa systems: a test of phenomenological melting criteria, *J. Chem. Phys.* **98**, 2319–2324 (1993).
- Sirota, E. B., Ou-Yang, H. D., Sinha, S. K. & Chaikin, P. M. Complete phase diagram of a charged colloidal system: a synchrotron X-ray scattering study, *Phys. Rev. Lett.* **62**, 1524–1527 (1989).
- Monovoukas, Y. & Gast, A. P. The experimental phase diagram of charged colloidal suspensions, *J. Colloid Interface Sci.* **128**, 533–548 (1989).
- Derjaguin, B. V. & Landau, L. Theory of the stability of strongly charged lyophobic sols and the adhesion of strongly charged particles in solutions of electrolytes, *Acta Physicochim. (USSR)* **14**, 633–662 (1941).
- Verwey, E. J. & Overbeek, J. Th. G. *Theory of the Stability of Lyophobic Colloids* (Elsevier, Amsterdam, 1948).
- Crocker, J. C. & Grier, D. G. Microscopic measurement of the pair interaction potential of charge-stabilized colloids, *Phys. Rev. Lett.* **73**, 352–355 (1994).
- Crocker, J. C. & Grier, D. G. Methods of digital video microscopy for colloidal studies, *J. Colloid Interface Sci.* **179**, 298–310 (1996).
- Crocker, J. C. & Grier, D. G. When like charges attract: the effects of geometrical confinement on long-range colloidal interactions, *Phys. Rev. Lett.* **77**, 1897–1900 (1996).
- Pailthorpe, B. A. & Russel, W. B. The retarded van der Waals interaction between spheres, *J. Colloid Interface Sci.* **89**, 563–566 (1982).
- Larsen, A. E. & Grier, D. G. Melting of metastable crystallites in charge-stabilized colloidal suspensions, *Phys. Rev. Lett.* **76**, 3862–3865 (1996).
- Happel, J. & Brenner, H. Viscous flow in multiparticle systems: Motion of spheres and a fluid in a cylindrical tube, *Am. Inst. Chem. Eng. J.* **3**, 506–513 (1957).
- Leamy, H. J. & Jackson, K. A. Roughness of the crystal-vapor interface, *J. Appl. Phys.* **42**, 2121–2127 (1971).
- Hoover, W. G. & Ree, F. H. Melting transition and communal entropy for hard spheres, *J. Chem. Phys.* **49**, 3609–3617 (1968).
- Hoover, W. G., Gray, S. G. & Johnson, K. W. Thermodynamic properties of the fluid and solid phases for inverse power potentials, *J. Chem. Phys.* **55**, 1128–1136 (1971).
- Curtin, W. A. Density functional theory of crystal-melt interfaces, *Phys. Rev. Lett.* **59**, 1228–1231 (1987).
- Curtin, W. A. Density-functional theory of the solid-liquid interface, *Phys. Rev. B* **39**, 6775–6791 (1989).
- Marr, D. W. & Gast, A. P. Planar density-functional approach to the solid-fluid interface of simple liquids, *Phys. Rev. E* **47**, 1212–1221 (1993).
- Kepler, G. M. & Fraden, S. Attractive potential between confined colloids at low ionic strength, *Phys. Rev. Lett.* **73**, 356–359 (1994).
- Carbajal-Tinoco, M. D., Castro-Román, F. Arauz-Lara, J. L. Static properties of confined colloidal suspensions, *Phys. Rev. E* **53**, 3745–3749 (1996).
- Sogami, I. Effective potential between charged spherical particles in dilute suspension, *Phys. Lett.* **96A**, 199–203 (1983).
- Sogami, I. & Ise, N. On the electrostatic interaction in macroionic solutions, *J. Chem. Phys.* **81**, 6320–6332 (1984).

Acknowledgements: We are grateful for conversations with D. Oxtoby, T. Palberg, S. Fraden, S. Rice, A. Marcus, T. Witten, C. Murray and, in particular, J. Crocker and M. Carbajal-Tinoco. This work was supported by the US National Science Foundation.

Correspondence and requests for materials should be addressed to D.G.G. (e-mail: grier@fakr.uchicago.edu).

Habitable moons around extrasolar giant planets

Darren M. Williams*, James F. Kasting†
& Richard A. Wade*

* Department of Astronomy and Astrophysics, The Pennsylvania State University, 525 Davey Laboratory, University Park, Pennsylvania 16802, USA

† Department of Geosciences, The Pennsylvania State University, 211 Deike Building, University Park, Pennsylvania 16802, USA

Possible planetary objects have now been discovered^{1–9} orbiting nine different main-sequence stars. These companion objects (some of which might actually be brown dwarfs) all have a mass at least half that of Jupiter, and are therefore unlikely to be hospitable to Earth-like life: jovian planets and brown dwarfs support neither a solid nor a liquid surface near which organisms might dwell. Here we argue that rocky moons orbiting these companions could be habitable if the planet–moon system orbits the parent star within the so-called ‘habitable zone’¹⁰, where life-supporting liquid water¹¹ could be present. The companions to the stars 16 Cygni B and 47 Ursae Majoris might satisfy this criterion. Such a moon would, however, need to be large enough (>0.12 Earth masses) to retain a substantial and long-lived atmosphere, and would also need to possess a strong magnetic field in order to prevent its atmosphere from being sputtered away by the constant bombardment of energetic ions from the planet’s magnetosphere.

Moons orbiting giant planets face additional problems that an isolated Earth-like planet would not encounter. At 1 AU from a solar-type star, all moons within a jovian planet’s Hill sphere (the limit for orbital stability) would become tidally locked within a few billion years of their formation. Hence, their rotations would last from a few days to a few months, which could lead to large diurnal temperature fluctuations. Moons of planets, or brown dwarfs, on highly eccentric orbits (for example, 70 Virginis B, HD114762 B, 16 Cygni Bb) might also experience large seasonal temperature fluctuations. Additionally, moons forming in a circumplanetary disk might receive different volatile endowments than would planets formed in a circumstellar disk. Also, moons orbiting jovian-type planets would presumably be embedded within the giant planet’s magnetosphere, where they would be constantly bombarded by energetic charged particles that could sputter away their atmospheres.

We consider the question of how moons are supplied with volatiles. The terrestrial planets are thought to have received most of their volatiles from bombardment by comets or carbonaceous asteroids¹². If giant planets form in the outer regions of circumstellar disks and then spiral inwards, as suggested by Lin *et al.*¹³, cometary bombardment would be a likely source of volatiles. (Whether giant planets can retain their moons as they migrate inward is an important, unanswered question.) Such a model has been proposed for the origin of Titan’s atmosphere¹⁴. Titan retained its atmosphere while the jovian moons did not, perhaps because impacts occurred at lower velocities in the smaller gravitational potential well of Saturn, which resulted in less atmospheric erosion. This suggests that moons orbiting Jupiter-sized planets or larger may have difficulty retaining cometary volatiles.

But moons formed in circumplanetary disks in the outer parts of a stellar nebula may have no need for additional volatiles. Indeed, the problem is just the opposite: many of these moons may be too volatile-rich. Ganymede and Callisto, which have densities of 1.8 g cm^{-3} and 1.5 g cm^{-3} , respectively, are at least half water-ice. If brought to 1 AU, the ice would melt and they would be covered with $\sim 1,000 \text{ km}$ of liquid water. This would not preclude habitability, but it would almost certainly preclude land-based life. Better candidates for producing Earth-analogue environments are moons

such as Io and Europa (densities of 3.5 g cm^{-3} and 3.0 g cm^{-3} , respectively) that are composed mostly of rock. Io has been devolatilized by tidally induced volcanism, but the difference between Europa and the two outermost galilean moons is considered to be primordial. Europa contains more rock and less ice because the inner region of the circum-jovian nebula was hotter¹⁵. If one assumes that extrasolar giant planets form in the same way, then clearly their inner moons are the most likely to be Earth-like.

A habitable moon must also be able to retain its volatiles for billions of years. Its effectiveness in doing so will depend on its mass, the flux of ionizing it receives from its star, and the charged particle flux it receives within the magnetosphere of its planet. For small moons with hot exospheres, thermal (Jeans) escape should be an important loss process. A moon having a Mars-like density of 3.9 g cm^{-3} , and an exosphere with its base at $1,000 \text{ km}$ and a temperature of $2,000 \text{ K}$, similar to Earth’s at solar maximum, could retain nitrogen and oxygen for over 4.5 Gyr if its mass were >0.07 times the mass of the Earth ($>0.07 M_{\oplus}$). Loss of O is not necessarily fatal to habitability because O_2 can be replaced by photosynthesis, and because large reservoirs of O exist in the form of water. Loss of N, though would be irreversible and could preclude the development of terrestrial-type life. Nitrogen is needed for metabolism by organisms, and N_2 is a non-toxic buffer gas that reduces inflammability of Earth-like ecosystems such as forests^{16,17}.

A second loss process for N, which is important on Mars, is dissociative recombination of N_2^+ (ref. 18). Mars loses nitrogen at a rate of $5 \times 10^5 \text{ atoms cm}^{-2} \text{ s}^{-1}$ (ref. 19). Scaling the martian loss rate to an N_2 -dominated atmosphere at 1 AU from the Sun implies a loss rate $\sim 4 \times 10^7 \text{ N atoms cm}^{-2}$, assuming that the ionization rate is proportional to the N_2 mixing ratio. This would remove $\sim 17\%$ of Earth’s N in 4.5 Gyr . So, moons with Earth-like N endowments would be only marginally affected, but moons with smaller N endowments might lose their atmospheres over time. This loss process would become negligible for moons with masses $\gtrsim 0.12 M_{\oplus}$ because the velocities of the non-thermal N atoms would be too slow to allow them to escape. Mars’ own mass is $\sim 0.1 M_{\oplus}$, which allows it to lose ^{14}N efficiently but not ^{15}N .

A potentially even greater threat to atmospheres of moons around giant planets is sputtering by charged particles trapped within the planet’s magnetosphere. Titan, which loses less than $1.2 \times 10^7 \text{ N atoms cm}^{-2} \text{ s}^{-1}$ as a consequence of sputtering²⁰ in the magnetosphere of Saturn (with mass ~ 0.3 times that of Jupiter, $\sim 0.3 M_J$), is only weakly affected. However, higher charged-particle densities could exist within the magnetospheres of more massive planets. A moon orbiting within Jupiter’s inner magnetosphere, for example, would receive an electron flux of $4 \times 10^8 \text{ cm}^{-2} \text{ s}^{-1}$ (ref. 21). By comparison, the solar-wind particle flux at Mars’s orbit is $4.8 \times 10^5 \text{ cm}^{-2} \text{ s}^{-1}$, which is thought to sputter away CO_2 and O at rates $\sim 2 \times 10^6 \text{ cm}^{-2} \text{ s}^{-1}$ (ref. 22). Thus, a non-magnetized moon in a jovian-like magnetosphere could lose N at a rate of $\sim 2 \times 10^9 \text{ cm}^{-2} \text{ s}^{-1}$. This would deplete Earth’s N_2 in only $\sim 5 \times 10^8 \text{ yr}$.

Atmospheric loss due to sputtering could effectively be eliminated, however, for moons with strong magnetic fields. Until recently, the smallest object in the Solar System known to possess an appreciable magnetic field was Mercury ($0.06 M_{\oplus}$). The remarkable new discovery by the Galileo spacecraft of a magnetosphere belonging to Ganymede^{23,24} ($0.03 M_{\oplus}$) suggests that at least some moons would be immune to this loss process.

Orbital constraints on habitability can be equally significant. Moons with captured rotations would have days which equal their orbital periods. The tidal locking radius for planets of different masses²⁵ is $r_t = 0.027(P_0 t/Q)^{1/6} M^{1/3}$. If we assume a primordial moon spin period P_0 of 15 h , a time t of 4.5 Gyr , and a tidal dissipation factor Q of 100 , then the 4.5-Gyr tidal locking radius (~ 96 times the radius of Jupiter, $\sim 96 R_J$) corresponds to a 116-day orbital period radius. Moons near this limit might experience large

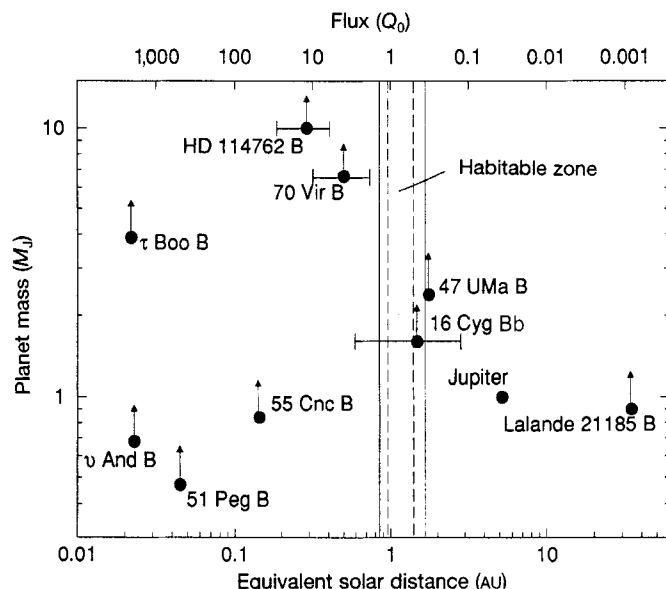


Figure 1 Locations of recently discovered or confirmed companions to nearby stars with respect to the habitable zone. The stars are: 51 Pegasi⁶, 70 Virginis⁵, 47 Ursae Majoris³, Lalande 21185 (ref. 4), 55 (p) Cancri⁵, τ Bootis⁵, ν Andromedae⁵, HD 114762 (refs 6, 9) and 16 Cygni Bb⁷ (a member of a triple star system—its planet is designated 16 Cygni Bb). Reported minimum planet masses are given relative to the mass of Jupiter (M_J). Planets are plotted according to the mean orbital flux, $\propto (1 - e^2)^{-1/2}$, they receive from their stars (upper x-axis) relative to Earth's solar constant $Q_0 = 1,370 \text{ W m}^{-2}$. The lower x-axis is the distance each planet would be from the Sun, in AU (astronomical units), to receive an equivalent flux. The range of stellar fluxes received by objects in highly eccentric orbits is indicated by solid horizontal lines crossing the circular markers. Conservative and non-conservative habitable zone limits¹⁰ are indicated by dashed lines and solid lines respectively.

diurnal temperature variations, but these moons, like Ganymede (15 R_J) and Callisto (22 R_J), would be covered by deep, high-heat-capacity oceans which would stabilize their climates over long rotations. Moons closer in (for example, Europa), having Earth-like atmospheres and surfaces, would experience substantial diurnal temperature swings in continental interiors, but these swings would be greatly moderated in coastal environments.

Moons belonging to objects in highly eccentric orbits are less likely to be habitable (Fig. 1). Objects in highly eccentric orbits are more likely to be brown dwarfs than planets⁶ (but see ref. 36 for an alternative interpretation). Heating from either brown dwarfs or planets would be negligible after the first billion years²⁷. The companion to the star 16 Cyg B has an orbital eccentricity, $e \approx 0.65$ (ref. 7), and the ratio of perihelion flux to aphelion flux is $[(1 + e)/(1 - e)]^2 \approx 22$. Land-based life would be hard-pressed to deal with such seasonal fluctuations, but the amplitudes of damaging seasonal extremes would be greatly reduced on moons possessing dense atmospheres. Hence, at least some moons on such eccentric orbits might still be suitable for life.

Our discussion of climate to this point has been based on short-term considerations. To remain habitable for an extended time period, an Earth-like planet or moon must have some mechanism to compensate for the gradual brightening of a star as it ages¹⁰. On Earth, such long-term climate regulation is thought to be provided by the carbonate-silicate cycle and its effect on atmospheric CO_2 (ref. 28). Cooler (warmer) climate conditions decrease (increase) the rate of silicate weathering and enhance (diminish) the concentration of the greenhouse gas, CO_2 . The presence of significant dry land is important in this feedback process because that is where

silicate weathering occurs. A second key factor is plate tectonics, which results in subduction and metamorphism of carbonate sediments and return of CO_2 to the atmosphere. The fact that Mars lost this capability early in its history may be one reason why it is so cold and desolate today²⁹.

A necessary, but not sufficient, condition for maintaining plate tectonics is the existence of a substantial heat flux from a planet's or moon's interior. Venus, which presumably has an Earth-like heat flux, does not have plate tectonics, perhaps because it has no water³⁰. Based on observations, the critical flux probably lies somewhere between that of Earth³¹, $\sim 70 \text{ erg cm}^{-2} \text{ s}^{-1}$, and Mars³², $\sim 30 \text{ erg cm}^{-2} \text{ s}^{-1}$.

To obtain an estimate of when enough radiogenic heat is available, we express the heat flux as $F_{\text{rad}} \propto R \rho e^{-\lambda t}$, where R and ρ are the moon's radius and density, and $\lambda = 1.5 \times 10^{-10} \text{ yr}^{-1}$ is the decay constant for ^{238}U . We further assume that Mars became tectonically inactive 2 Gyr ago³³, which implies a critical heat flow of $\sim 40 \text{ erg cm}^{-2} \text{ s}^{-1}$. For a density similar to that of Io, this yields a lower limit of $\sim 0.23 M_{\oplus}$ for a moon capable of sustaining plate tectonics.

This 'tectonic' lower mass limit might be circumvented if the innermost moons experienced substantial tidal heating, as is true for Io and, to a lesser extent, Europa. The rate of tidal heating is $F_{\text{tid}} = (9/19) \rho^2 n^5 R^5 e^2 (\mu Q)^{-1}$, where n is the moon's mean motion about the planet, and μ is the moon's rigidity ($6.5 \times 10^{11} \text{ dyn cm}^{-2}$ for Io)³⁴. Io's eccentricity is maintained at a finite value (0.004), despite tidal damping, by a 4:2:1 mean-motion resonance with Europa and Ganymede. If such resonances arise naturally during formation, as seems likely³⁵, similar relationships might occur among the moons of extrasolar giant planets. Io's tidal surface heating is calculated to be $\sim 41 \text{ erg cm}^{-2} \text{ s}^{-1}$, perhaps enough to drive plate tectonics on a more Earth-like moon. Alternatively, point volcanism of the sort that actually occurs on Io might be sufficient to recycle carbonate rocks and maintain an active carbonate-silicate cycle. So, close-in moons with masses $< 0.23 M_{\oplus}$ could conceivably remain habitable for long time periods. Io itself would not fulfill this criterion because it is too small to retain its volatiles.

Some moons around extrasolar giant planets might be habitable, but only if they are of sufficient size. A $0.12 M_{\oplus}$ moon in an Io-like orbital resonance, and possessing a Ganymede-like magnetic field, could conceivably remain habitable for billions of years, given the right amount of stellar insolation. All this suggests that the systems belonging to 47 UMa and 16 Cyg B should be considered as possible abodes for extraterrestrial life. $\square$

Received 18 July; accepted 20 November 1996.

- Mayor, M. & Queloz, D. *Nature* **378**, 355–359 (1995).
- Marcy, G. W. & Butler, R. P. *Astrophys. J.* **464**, L147–L151 (1996).
- Butler, R. P. & Marcy, G. W. *Astrophys. J.* **464**, L153–L156 (1996).
- Gatewood, G. *Bull. Am. Astron. Soc.* **28**, 885 (1996).
- Butler, R. P., Marcy, G. W., Williams, E., Hauser, H. & Shirts, P. *Astrophys. J.* (submitted).
- Latham, D. W., Mazeh, T., Stefanik, R. P., Mayor, M. & Burki, G. *Nature* **339**, 38–40 (1989).
- Cochran, W. D., Hatzes, A. P., Butler, R. P. & Marcy, G. W. *Bull. Am. Astron. Soc.* **28**, 1111 (1996).
- Marcy, G. W., Butler, R. P., Williams, E., Bildsten, L. & Graham, J. R. *Astrophys. J.* (submitted).
- Cochran, W. D., Hatzes, A. P. & Hancock, T. J. *Astrophys. J.* **380**, L35–L38 (1991).
- Kasting, J. F., Whitmire, D. P. & Reynolds, R. T. *Icarus* **101**, 1–21 (1993).
- Brack, A. *Orig. Life* **23**, 3–10 (1993).
- Chyba, C. F. *Nature* **343**, 129–133 (1990).
- Lin, D., Bodenheimer, P. & Richardson, D. C. *Nature* **380**, 606–607 (1996).
- Griffith, C. A. & Zahnle, K. J. *Geophys. Res.* **100**, 16907–16922 (1995).
- Prinn, R. G. & Fegley, B. *Astrophys. J.* **249**, 308–317 (1981).
- Watson, A., Lovelock, J. E. & Margulis, L. *Biosystems* **10**, 293–298 (1978).
- McKay, C. P., Toon, O. B. & Kasting, J. F. *Nature* **352**, 489–496 (1991).
- McElroy, M. B. *Science* **175**, 443–445 (1972).
- Fox, J. L. *J. Geophys. Res.* **98**, 3297–3310 (1993).
- Strobel, D. F., Summers, M. E. & Zhu, X. *Icarus* **100**, 512–526 (1992).
- Van Allen, J. A. in *Jupiter* (ed. Gehrels, T.) (Univ. Arizona Press, Tucson, 1976).
- Kass, D. M. & Yung, Y. L. *Science* **268**, 697–699 (1995).
- Kivelson, M. G. et al. *Nature* **384**, 537–541 (1996).
- Gurnett, D. A., Kurtin, W. S., Rowe, A., Bolton, S. J. & Kennel, C. F. *Nature* **384**, 535–537 (1996).
- Peale, S. J. in *Planetary Satellites* (ed. Burns, J. A.) 87–112 (Univ. Arizona Press, Tucson, 1977).
- Boss, A. P. *Nature* **379**, 397–398 (1996).
- Saumon, D. et al. *Astrophys. J.* **460**, 993–1018 (1996).
- Walker, J. C. G., Hays, P. B. & Kasting, J. F. *J. Geophys. Res.* **86**, 9776–9782 (1981).
- Kasting, J. F., Toon, O. B. & Pollack, J. B. *Sci. Am.* **256**, 90–97 (1988).

30. Turcotte, D. L. *J. Geophys. Res.* **101**, 4765–4773 (1996).
31. Turcotte, D. L. & Schubert, G. *Geodynamics: Applications of Continuum Physics to Geologic Problems* 137–139 Wiley, New York, 1982).
32. Solomon, S. C. & Head, J. W. *J. Geophys. Res.* **95**, 11073–11083 (1990).
33. Francis, P. W. & Wood, C. A. *J. Geophys. Res.* **87**, 9881–9889 (1982).
34. Peale, S. J., Cassen, P. & Reynolds, R. T. *Science* **203**, 892–894 (1979).
35. Roy, A. E. *Orbital Motion* 3rd edn 239–279 (Hilger, Philadelphia, 1988).
36. Rasio, F. A. & Ford, E. B. *Science* **274**, 954–956 (1996).

Acknowledgements: We thank K. Zahnle and R. Lorenz for discussions. D.W. was supported by a NASA graduate student research fellowship (1995); J.K. was supported by the NASA Exobiology Program.

Correspondence and requests for materials should be addressed to D.W. (e-mail: dwilliams@astro.psu.edu).

Evidence for polaronic supercarriers in the copper oxide superconductors $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$

Guo-meng Zhao, M. B. Hunt, H. Keller & K. A. Müller

Physik-Institut der Universität Zürich, CH-8057 Zürich, Switzerland

High-temperature superconductivity is known to involve the pairing of charge carriers, but the precise nature of these carriers and the mechanism of their pairing remain unclear. The copper oxides are known to exhibit a strong Jahn–Teller effect — in which spontaneous lattice distortions remove the degeneracy of the electronic ground state — and it has been suggested that the charge carriers are Jahn–Teller polarons (bare charge carriers accompanied by local lattice distortions). In fact, the demonstration¹ that a strong Jahn–Teller effect can lead to the formation of such polarons led to the original discovery of high-temperature superconductivity². Still, direct evidence that Jahn–Teller polarons exist in the superconducting state of the copper oxides has been lacking, although some indirect evidence comes from their recent discovery³ in the structurally similar but non-superconducting manganite $\text{La}_{1-x}\text{Ca}_x\text{MnO}_3$. Here we report the results of magnetization and thermal expansion measurements on samples of the copper oxide superconductor $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$ which characterize the oxygen-isotope effects on the carrier density and on the in-plane penetration depth. We find a negligible isotope effect on the former, but a large effect on the latter. Specific quantitative features of the results show that polaronic charge carriers exist and condense into Cooper pairs in the copper oxide superconductors.

Recently there has been increasing evidence that polaronic charge carriers are present in the normal state of the layered copper oxides^{4,5}. However, it is not clear whether these normal-state polaronic carriers condense into Cooper pairs. To show that this occurs, it is necessary to demonstrate that the effective supercarrier mass along the CuO_2 planes (m_{ab}^*) depends strongly on ionic mass M in these layered copper oxide superconductors. In conventional superconductors, only the ‘bare’ carriers condense into the supercarriers, and the supercarrier mass is essentially independent of M . In the polaronic model⁶, the effective mass of polarons depends strongly on M . When the polarons condense into supercarriers (by whatever pairing mechanism), the effective supercarrier mass will also depend on M .

Several groups^{7–11} have observed an anomalously large oxygen-isotope effect on the superconducting transition temperature T_c in underdoped copper oxides, and all of them^{8–11} noticed that there is an isotope-dependence of the diamagnetic signal. Zhao *et al.*¹² have studied this effect and tried to interpret it as due to the isotope-mass dependence of the average supercarrier mass m^* ($= [(m_{ab}^*)^2 m_c^*]^{1/3}$), where m_{ab}^* and m_c^* are the effective supercarrier mass along and

perpendicular to the CuO_2 plane, respectively). However their preliminary data analysis could not determine whether, and how strongly, m_{ab}^* should depend on the oxygen mass. A quantitative determination of the oxygen-isotope effect on m_{ab}^* is crucial to the understanding of the physics of high-temperature superconductivity.

Because the magnetic penetration depth $\lambda(0)$ is proportional to $\sqrt{m^*/n_s}$, where n_s is the supercarrier density, then

$$\Delta m^*/m^* = 2\Delta\lambda(0)/\lambda(0) + \Delta n_s/n_s, \quad (1)$$

where Δ means any small change of a quantity upon isotope substitution. Thus the isotope dependence of m^* can be determined if the isotope dependences of $\lambda(0)$ and of n_s can be independently measured.

The isotope dependence of $\lambda(0)$ can be determined from that of the Meissner fraction $f(0)$ which, for decoupled and fine-grained samples, depends on the penetration depth $\lambda(0)$ and on the average grain radius R , as seen from the Shoenberg formula for spherical grains¹³:

$$f(T) = \frac{3}{2} \left[1 - 3 \left(\frac{\lambda(T)}{R} \right) \coth \left(\frac{R}{\lambda(T)} \right) + 3 \left(\frac{\lambda(T)}{R} \right)^2 \right] \quad (2)$$

where $\lambda(T) = [(\lambda_{ab}(T))^2 \lambda_c(T)]^{1/3}$ for layered compounds (where $\lambda_{ab}(T)$ and $\lambda_c(T)$ are the penetration depths along and perpendicular to the CuO_2 plane, respectively)¹⁴. From equation (2), it may be seen that a change in $\lambda(0)$ will lead to a change in $f(0)$, so the isotope dependence of $\lambda(0)$ can be determined from the isotope dependence of $f(0)$.

The isotope dependence of n_s should be equal to the isotope dependence of the normal carrier density n in clean superconductors, as is the case in the underdoped copper oxide superconductors¹⁵. In general, the carrier density n cannot be measured very precisely by any direct experimental methods, so it is very difficult to determine the isotope dependence of n . Nevertheless, we note that the structural phase transition temperature T_s (from the tetragonal to orthorhombic phase) in the $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$ system is very sensitive to n . This implies that a very small difference in n will lead to a huge difference in T_s . Thus the isotope dependence of n can be measured precisely if one can accurately determine the isotope shift of T_s . It has also been found that there is an anomaly at T_s in the thermal-expansion coefficient of this system^{16,17}. Therefore the isotope shift of T_s can be precisely determined by a thermal-expansion measurement.

Samples of $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$ were prepared by conventional solid-state reaction using La_2O_3 (99.99%), SrCO_3 (99.999%) and CuO (99.999%). For the underdoped samples, we used the sample preparation procedures described in ref. 12. To obtain samples with small grains and sufficient porosity for the isotope experiments, we reground the samples thoroughly, pressed them into pellets, and annealed them in air at 900 °C for 12 h. Each pellet was broken in half, and the halves were then subjected to ^{16}O and ^{18}O isotope diffusion. The diffusion was carried out for 40 h at 900 °C and oxygen pressure of ~ 1 bar. The cooling time to room temperature was 4 h. The oxygen isotope enrichment was determined from the weight changes of the ^{16}O and ^{18}O samples. The ^{18}O samples had 85–90% ^{18}O and ~ 15 –10% ^{16}O . Back-exchange was carried out at 850 °C for 8 h in air. For the optimally doped samples ($x = 0.15$), the well-mixed powders were calcined for 15 h at 1,000 °C in air. The sintering was performed for 15 h at 1,000 °C in air. Before oxygen isotope diffusion, the samples were ground for about 40 min to ensure that the grain size was as fine as possible. The powder samples were then subject to ^{16}O and ^{18}O isotope diffusion at 900 °C for 20 h. The ^{18}O samples had more than 95% ^{18}O .

Magnetization was measured with a Quantum Design SQUID magnetometer. The Meissner effect was measured in a magnetic field of 1 mT after the samples had been cooled from the normal state in the field. The magnetic field remained unchanged during

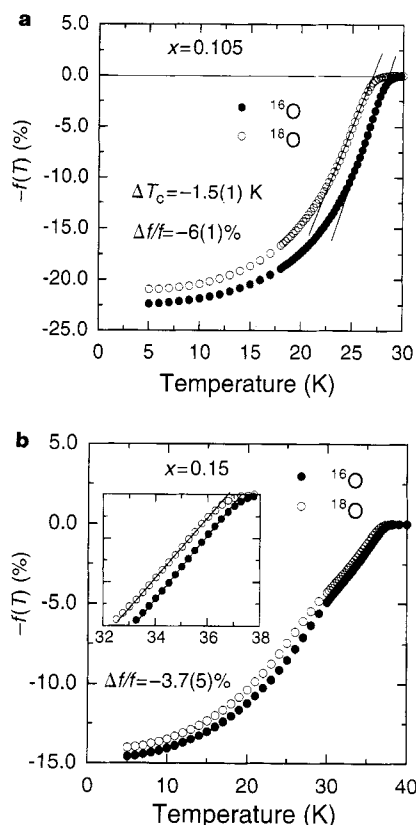


Figure 1 Temperature dependence of the Meissner effect for ^{16}O and ^{18}O samples of $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$ with $x = 0.105$ (**a**), and $x = 0.15$ (**b**). Note that for $x = 0.15$ there is a small 'bump' in $f(T)$ due to a change in the effective R caused by the field-induced decoupling of some weak links.

each series of measurements. For the $x = 0.15$ samples, we suspended the powders in paraffin to prevent them from moving during magnetization measurements. This technique is essential for the success of this isotope experiment as pelletized samples have too strong intergrain flux trapping at this composition. The thermal expansion was measured on the underdoped samples using a capacitance dilatometer with a length resolution of $\sim 0.1 \text{ \AA}$. Data were collected during warming and subsequent cooling at the same constant rate of 3 mK s^{-1} .

Figure 1 shows the temperature dependence of the Meissner effect for the ^{16}O and ^{18}O samples of $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$ with $x = 0.105$ (Fig. 1a), and $x = 0.15$ (Fig. 1b). For $x = 0.105$, the T_c of the ^{18}O sample is lower than that of the ^{16}O sample by $\sim 1.5 \text{ K}$ and the low-temperature Meissner fraction of the ^{18}O sample is reduced by $6(1)\%$ relative to the ^{16}O sample (note that $6(1)\%$ means $6(\pm 1)\%$). This is in agreement with the result reported in ref. 12, indicating the excellent reproducibility of these isotope experiments. For the optimally doped samples ($x = 0.15$), the T_c of the ^{18}O sample is lower than that of the ^{16}O sample by $0.52(4) \text{ K}$, and the low-temperature Meissner fraction of the ^{18}O sample is reduced by $3.7(5)\%$ relative to the ^{16}O sample. Both T_c and the isotope shift are in good agreement with those reported by Franck *et al.*¹⁰. The broad transition in the $x = 0.15$ samples is due to a small average grain radius ($2R \approx 1 \mu\text{m}$)⁹. We have also shown that there is a small flux trapping in our samples ($<15\%$ of the Meissner effect), so it will have a negligible effect on the isotope dependence of the Meissner fraction¹².

It is essential to show that the observed isotope effects are intrinsic. To do this, we took one set of ^{16}O and ^{18}O samples of $\text{La}_{1.895}\text{Sr}_{0.105}\text{CuO}_4$, and annealed them together in an atmosphere of $0.2 \text{ bar } ^{16}\text{O}_2$ at 850°C for 8 h. After annealing, the T_c and Meissner

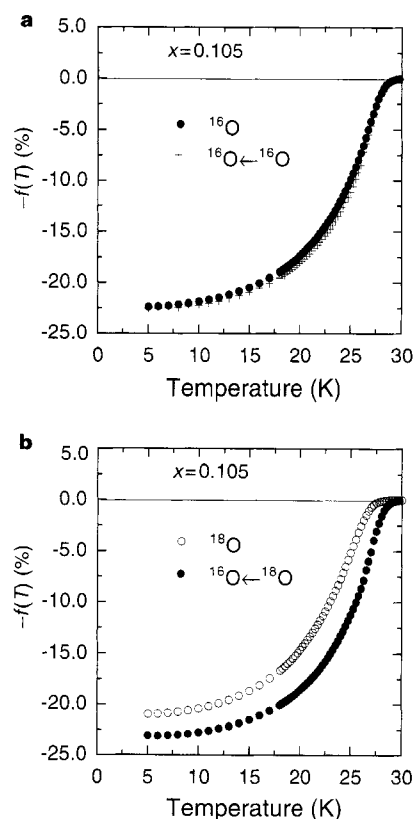


Figure 2 Temperature dependence of the Meissner effect for one set of ^{16}O and ^{18}O samples of $\text{La}_{1.895}\text{Sr}_{0.105}\text{CuO}_4$ before and after isotope back-exchange: from ^{16}O to ^{18}O (**a**); from ^{18}O to ^{16}O (**b**).

effect for the ^{16}O sample were nearly unchanged (Fig. 2a). This implies that such a treatment has a negligible effect on the carrier concentration and grain size of the sample. On the other hand, the T_c and Meissner effect for the ^{18}O sample changed dramatically (Fig. 2b) due to the back substitution of ^{16}O for ^{18}O ; the T_c of the sample is reduced by $\sim 1.5 \text{ K}$, and the low-temperature Meissner fraction decreases $8(1)\%$. These results show clearly that the oxygen-isotope substitution indeed leads to substantial reductions in both T_c and Meissner fraction.

In Fig. 3a we show the linear thermal expansion coefficient $\beta(T)$ for another set of ^{16}O and ^{18}O samples of $\text{La}_{1.895}\text{Sr}_{0.105}\text{CuO}_4$, which shows the same oxygen-isotope effects as the set shown in Fig. 1a. There is an anomaly in $\beta(T)$ at a temperature of $\sim 265 \text{ K}$ (marked by an arrow in Fig. 3a), which corresponds to the tetragonal–orthorhombic transition temperature T_s (ref. 17). It is clear that the two curves nearly overlap, implying that the two isotope samples have the same T_s within the uncertainty of our measurements ($\leq 0.3 \text{ K}$). In Fig. 3b we show the dependence of T_s on the Sr content x , as determined by further thermal expansion measurements. A linear fit to the data with $T_s = A - Bx$ gives $A = 569 \text{ K}$ and $B = 2,900 \text{ K}$. This relation implies that $T_s = 569 \text{ K}$ for $x = 0$, and $T_s = 0$ for $x = 0.20$, which is in fair agreement with reported results^{18,19}. It has also been shown by Dabrowski *et al.*¹⁹ that T_s is mainly determined by the hole concentration, although the mean ionic radius at the cation site has a small effect on T_s (for example, the composition x_c at which $T_s = 0$ is ~ 0.2 in $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$, and ~ 0.25 in $\text{La}_{2-x}\text{Ca}_x\text{CuO}_4$). Then, with $n = x$, we can write $dT_s/dn \approx dT_s/dx = B$. Using $\Delta T_s \leq 0.3 \text{ K}$ and $n = 0.105$, we obtain $|\Delta n/n| = |(1/nB)\Delta T_s| \leq 0.1\%$. Thus, there is a negligible oxygen-isotope effect on n . In addition, we have shown that the ratio of the normal-state susceptibilities (including a small Curie

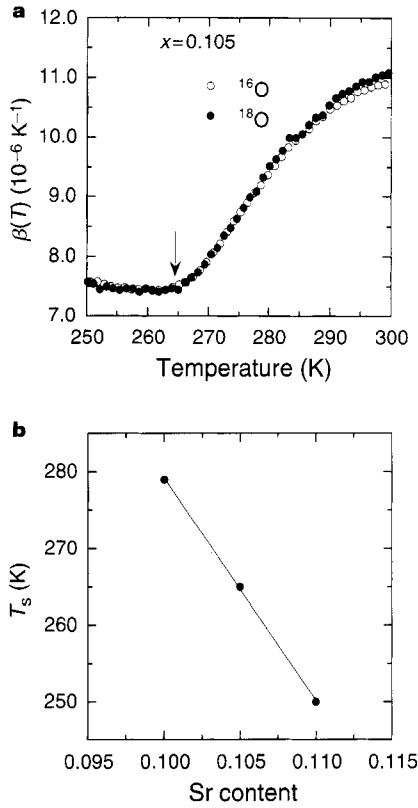


Figure 3 The linear thermal expansion coefficient $\beta(T)$ for the ^{16}O and ^{18}O samples of $\text{La}_{1.895}\text{Sr}_{0.105}\text{CuO}_4$ (**a**), and the dependence of the tetragonal-orthorhombic transition temperature T_s on the Sr content x (**b**).

term contributed by some localized carriers) for the two isotope samples is temperature-independent down to a temperature near T_c . This implies that the two isotope samples must have the same mobile carrier concentration near T_c even if charge localization occurs. Fiory *et al.*²⁰ have shown that the carrier concentration in slightly underdoped $\text{YBa}_2\text{Cu}_3\text{O}_{7-y}$ is nearly the same in the normal and superconducting states, as expected for a clean superconductor. We would expect the same result in the clean one-layer superconductor $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$. Thus our results imply that there is a negligible oxygen-isotope effect on the supercarrier density n_s .

We now consider the determination of the oxygen-mass dependence of the penetration depth from equation (2) and the measured oxygen-isotope effect on $f(0)$. From equation (2), we obtain

$$\Delta\lambda(0)/\lambda(0) = C(R)\Delta f(0)/f(0) \quad (3)$$

where $C(R) = [f(0)/\lambda(0)](\partial\lambda/\partial f)$, which implicitly depends on $f(0)$ (for example, for $f(0) = 22\%$, $C(R) = -0.63$; $f(0) = 14\%$, $C(R) = -0.57$). From equation (3), and using measured values of $f(0) = 22\%$ and $\Delta f(0)/f(0) = -6(1)\%$ for $x = 0.105$, we find $\Delta\lambda(0)/\lambda(0) = 3.8(5)\%$. Similarly, for $x = 0.15$, using $f(0) = 14\%$ and $\Delta f(0)/f(0) = 3.7(5)\%$, we find $\Delta\lambda(0)/\lambda(0) = 2.1(3)\%$. Shibauchi *et al.*²¹ showed that, due to the intrinsic Josephson coupling between the CuO_2 planes, the penetration depth along the c -direction $\lambda_c(0) \propto \sqrt{(\rho_c/T_c)}$ in $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$, where ρ_c is the resistivity along the c -direction. The resistivity is independent of the isotope mass in the case of the polaronic hopping conductivity observed in this system⁵. Using $\lambda(0) = ([(\lambda_{ab}(0))^2\lambda_c(0)]^{1/3}$ we find

$$\Delta\lambda_{ab}(0)/\lambda_{ab}(0) = \frac{3}{2}[\Delta\lambda(0)/\lambda(0)] + \frac{1}{4}(\Delta T_c/T_c) \quad (4)$$

Substituting $\Delta\lambda(0)/\lambda(0) = 3.8(5)\%$ and $\Delta T_c/T_c = -5.4(4)\%$ for $x = 0.105$, into equation (4), we get $\Delta\lambda_{ab}(0)/\lambda_{ab}(0) = 4.4(5)\%$.

Then $\Delta m_{ab}^*/m_{ab}^* = 2\Delta\lambda_{ab}(0)/\lambda_{ab}(0) = 9(1)\%$ given that $\Delta n_s/n_s = 0$. If we define the exponent of the oxygen-isotope effect on m_{ab}^* as $\alpha_m^o = -\text{dln} m_{ab}^*/\text{dln} M_o$, then $\alpha_m^o = -0.8(1)$ for $x = 0.105$. In the same way, we obtain $\Delta m_{ab}^*/m_{ab}^* = 5.6(5)\%$ and $\alpha_m^o = 0.46(5)$ for $x = 0.15$.

We now discuss the isotope dependence of m_{ab}^* on the basis of the polaron model⁶. In this model, the effective mass of polarons (m_{ab}^p) is given as

$$m_{ab}^p \propto \exp(\gamma E_b/\hbar\omega) \quad (5)$$

where E_b is the binding energy of polarons, and is independent of the isotope mass, γ is a dimensionless constant ($0 \leq \gamma \leq 1$) which increases with increasing E_b/t (where t is the bare hopping integral) and ω is the characteristic frequency of the optical phonons. From equation (5), the total exponent of the isotope effect on m_{ab}^p is

$$\alpha_{mp}^i = -\text{dln} m_{ab}^p/\text{dln} M_o = -0.5\gamma E_b/\hbar\omega \quad (6)$$

Furthermore, if we assume that the polaronic carriers condense into Cooper pairs, then $m_{ab}^* = m_{ab}^p(1 + \lambda_{cb})$, where the coupling constant $\lambda_{cb} = N^p(0)V$ with $N^p(0)$ being the density of states of the polaronic conduction band ($\propto m_{ab}^p$) and V being the pairing potential which could arise from phonons and/or other bosonic excitations. It is well established that the coupling in the hole-doped cuprates is very strong ($\lambda_{cb} \geq 3$), as seen from the large reduced gap ($2\Delta_{\text{max}}/k_B T_c = 8$) measured for optimally doped $\text{YBa}_2\text{Cu}_3\text{O}_7$, $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_8$ and $\text{La}_{1.83}\text{Sr}_{0.17}\text{CuO}_4$ compounds^{22,23}. For $\lambda_{cb} \gg 1$, $m_{ab}^* \propto (m_{ab}^p)^2$ so that $\Delta m_{ab}^*/m_{ab}^* \approx \frac{1}{2}\Delta m_{ab}^p/m_{ab}^p$ or $\alpha_{mp}^i = \frac{1}{2}\alpha_{mp}^i$. Moreover, one would expect that $\alpha_{mp}^i = 2\alpha_m^o$ because the oxygen- and copper-isotope effects on T_c have the same exponent in underdoped $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$ (ref. 10), and because there should be no contribution to the isotope effect from La and Sr atoms. Then we have $\alpha_{mp}^i = \alpha_m^o$, leading to

$$\alpha_m^o = -0.5\gamma E_b/\hbar\omega \quad (7)$$

From equations (5) and (7), we find that the effective mass of polarons is enhanced by a factor $\exp(\gamma E_b/\hbar\omega) = \exp(-2\alpha_m^o)$ with respect to the bare mass. For $x = 0.105$, this enhancement factor is intermediate (~ 5.0), and decreases further towards the optimally doped regime (~ 2.5). As the coupling constant should be nearly doping-independent owing to the doping-independence of the superconducting gap^{24,25}, the supercarrier mass m_{ab}^* for the optimally-doped sample will be about half that for the $x = 0.105$ sample. This implies that m_{ab}^*/n_s for the optimally-doped sample is about one-third that for the $x = 0.105$ sample (because n_s for $x = 0.15$ is ~ 1.5 times larger than that for $x = 0.105$), in excellent agreement with the muon spin rotation measurement¹⁰. Moreover, from equation (7) and the measured values of α_m^o for $x = 0.105$ and 0.15 , we see the polaron binding energy E_b for $x = 0.15$ is about half that for $x = 0.105$, in fair agreement with the optical measurements in this system ($E_b = 0.12$ eV for $x = 0.10$, and $E_b = 0.06$ eV for $x = 0.15$)⁵. The above consistency leads us to conclude that polaronic charge carriers exist and condense into Cooper pairs in high-temperature superconductors. The mechanism for the formation of polarons may arise from a strong Jahn–Teller effect, as is the case in manganites³, which would be consistent with the theory proposed by Höck *et al.*¹ □

Received 19 August; accepted 20 November 1996.

- Höck, K.-H., Nickisch, H. & Thomas, H. *Helv. Phys. Acta* **56**, 237–243 (1983).
- Bednorz, J. G. & Müller, K. A. *Z. Phys. B* **64**, 189–193 (1986).
- Zho, G. M., Conder, K., Keller, H. & Müller, K. A. *Nature* **381**, 676–678 (1996).
- Proc. Int. Workshop on Anharmonic Properties of High- T_c Cuprates (eds Mihailovic, D., Ruani, G., Kaldis, E. & Müller, K. A.) 118–146 (World Scientific, Singapore, 1994).
- Bi, X. X. & Eklund, P. C. *Phys. Rev. Lett.* **70**, 2625–2628 (1993).
- Alexandrov, A. S. & Mott, N. F. *Int. J. Mod. Phys. B* **8**, 2075–2109 (1994).
- Crawford, M. K., Kunchur, M. N., Farneth, W. E., McCarron, E. M. & Poon, S. J. *Phys. Rev. B* **41**, 282–287 (1990).
- Crawford, M. K., Farneth, W. E., McCarron, E. M., Harlow, R. L. & Moudren, A. H. *Science* **250**, 1390–1394 (1990).
- Ronay, M., Frisch, M. A. & McGuire, T. R. *Phys. Rev. B* **45**, 355–360 (1992).
- Franck, J. P., Harker, S. & Brewer, J. H. *Phys. Rev. Lett.* **71**, 283–286 (1993).

11. Bornemann, H. J. & Morris, D. E. *Phys. Rev. B* **44**, 5322–5325 (1991).
12. Zhao, G. M., Singh, K. K., Sinha, A. P. B. & Morris, D. E. *Phys. Rev. B* **52**, 6840–6844 (1995).
13. Shoenberg, D. *Proc. R. Soc. Lond. A* **175**, 49–70 (1940).
14. Kogan, V. G., Fang, M. M. & Mitra, S. *Phys. Rev. B* **38**, 11958–11961 (1988).
15. Uemura, Y. J. *et al. Phys. Rev. Lett.* **62**, 2317–2320 (1989).
16. Sarrao, J. L. *et al. Phys. Rev. B* **50**, 13125–13131 (1994).
17. Okajima, Y., Hashimoto, S. & Yamaya, K. *Physica C* **235–240**, 1317–1318 (1994).
18. Kato, M., Maeno, Y. & Fujita, T. *Physica C* **152**, 116–120 (1988).
19. Dabrowski, B. *et al. Phys. Rev. Lett.* **76**, 1348–1351 (1996).
20. Fiory, A. T. *et al. Phys. Rev. Lett.* **65**, 3441–3444 (1990).
21. Shibauchi, T. *et al. Phys. Rev. Lett.* **72**, 2263–2266 (1994).
22. Ekino, T. M., Fujii, H. & Akimitsu, J. *Physica C* **235–240**, 1899–1900 (1994).
23. Chen, X. K., Irwin, J. C., Trodahl, H. J., Kimura, T. & Kishio, K. *Phys. Rev. Lett.* **73**, 3290–3293 (1994).
24. Ding, H. *et al. Nature* **382**, 51–54 (1996).
25. Loram, J. W., Mirza, K. A., Wade, J. M., Cooper, J. R. & Liang, W. Y. *Physica C* **235–240**, 134–137 (1994).

Acknowledgements: We thank W. Paul for use of the dilatometer at the ABB Research Center. This work was supported by the Swiss National Science Foundation.

Correspondence and requests for materials should be addressed to H.K. (e-mail: keller@physik-rzu.unizh.ch).

Conversion of light energy to proton potential in liposomes by artificial photosynthetic reaction centres

Gali Steinberg-Yfrach, Paul A. Liddell, Su-Chun Hung, Ana L. Moore, Devens Gust & Thomas A. Moore

Center for The Study of Early Events in Photosynthesis and
Department of Chemistry & Biochemistry, Arizona State University,
Tempe, Arizona 85287-1604, USA

During photosynthesis, photoinduced electron transport across membranes is carried out by pigment molecules organized into reaction centres by membrane-spanning proteins. The resulting transmembrane electrochemical potential is then coupled to the movement of protons across the membrane¹. Photoinduced electron transport followed by thermal electron transfer, leading to charge separation over distances of 8 nm, has been demonstrated in artificial mimics of the photosynthetic reaction centre comprising covalently linked electron donors and acceptors^{2–10}. Here we report the assembly of an artificial mimic of the photosynthetic apparatus which transports protons across a lipid bilayer when illuminated. Our model reaction centre is a molecular 'triad', consisting of an electron donor and acceptor linked to a photo-sensitive porphyrin group. This triad is incorporated into the bilayer of a liposome. When excited, it establishes a reduction potential near the outer surface of the bilayer and an oxidation potential near its inner surface. In response to this redox potential gradient, a freely diffusing quinone molecule alternates between its oxidized and reduced forms to ferry protons across the bilayer with an overall quantum yield of 0.004, creating a pH gradient between the inside and outside of the liposome.

Molecular triad **1** (Fig. 1 bottom) was prepared by linking a synthetic tetraarylporphyrin (P) to both a naphthoquinone moiety fused to a norbornene system bearing a carboxylic acid group (Q) and a carotenoid polyene (C)¹¹. On excitation, the triad (dissolved in various solvents) undergoes photoinduced electron transfer from the excited singlet state of the porphyrin moiety to yield the intermediate charge-separated species $C^+-P^+-Q^-$ with a quantum yield of ~ 1 . Subsequent electron transfer from the carotenoid to the porphyrin radical cation competes with charge recombination to give C^+-P-Q^- with a quantum yield up to 0.15 (ref. 11).

Liposomes were prepared by standard reverse-phase evaporation

procedures¹² from a lipid mixture that contained the lipid-soluble 2,5-diphenylbenzoquinone (Q_s ; Fig. 1 bottom) as the proton shuttle. Vectorial electron and proton transport requires asymmetric insertion of the molecular triad artificial reaction centre into the liposomal bilayer membrane. This was accomplished by injection of a tetrahydrofuran solution of **1** into an aqueous solution containing liposomes, a procedure which delivers **1** to the outside surface of the bilayer. A preference for the directional order indicated in Fig. 1 (top) is expected because insertion of the lipophilic carotenoid of **1** into the bilayer avoids the energetic cost of moving the polar, carboxylate-bearing quinone through the low-dielectric medium.

Laser flash photolysis experiments on liposomes containing **1** are

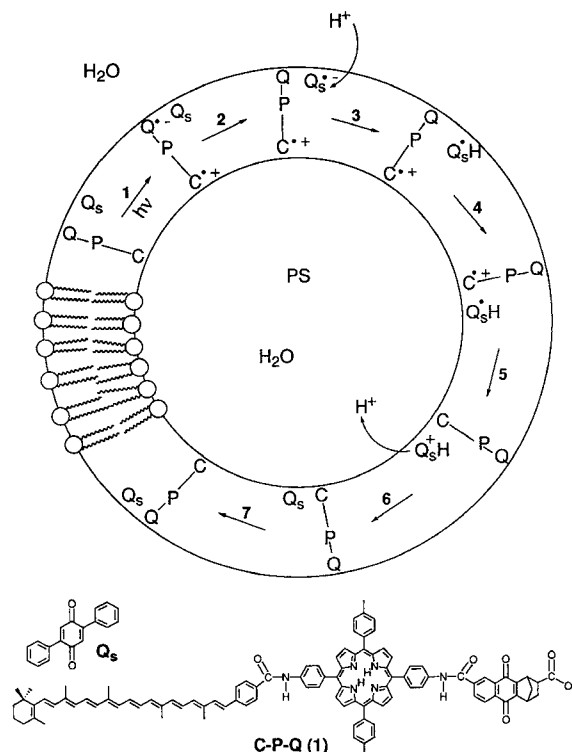


Figure 1 Bottom, Structures of triad **1** and 2,5-diphenylbenzoquinone (Q_s); top, schematic diagram illustrating the orientation of triad **1** in a liposome, with its quinone near the interface between the bulk aqueous phase and the bilayer lipid membrane. Steps 1–7 illustrate the proposed elementary processes involved in proton translocation across the liposomal bilayer. Reverse-phase evaporation (RPE) liposomes were prepared by standard procedures using 5 mg of a 2:3 molar ratio of L- α -phosphatidylserine (brain, $\sim 60\%$ 1-stearoyl-2-oleoylphosphatidylserine) and dioleoylphosphatidylcholine¹². The solution also contained 0.05 M KCl and pyraninetrisulphonate (PS), a water-soluble dye which indicates the pH by the ratio of the amplitude of the fluorescence excitation spectrum at 406 nm to that at 456 nm (I_{406}/I_{456}) (ref. 18). For certain experiments, the lipid mixture also contained 0.4 mg of lipid-soluble Q_s as the proton shuttle. The midpoint potential for the one-electron reduction of Q_s is more positive by ~ 0.6 V than that of the quinone component of **1** (ref. 19). Triad **1** was added to a stirred solution of liposomes by injection of 30–40 μ l of a 7.7×10^{-5} M solution of **1** in tetrahydrofuran. The liposomes were then purified by gel chromatography (Sephadex G-100; elution with a 0.05 M KCl solution). Quantitative analysis²⁰ of the organic phosphorus in these liposomes indicated that they typically contain $\sim 5 \times 10^{17}$ molecules of lipid per ml. RPE liposomes formed in this way have $\sim 8 \times 10^4$ molecules lipid per liposome (therefore, $\sim 5 \times 10^{12}$ liposomes per ml), an average diameter of 100 nm and an average volume of 5×10^{-18} l (ref. 12). Based on the amount of dye and **1** extracted from a typical liposome preparation, there are ~ 40 molecules of **1** in the bilayer and ~ 50 molecules of PS in the interior aqueous phase of each liposome.

consistent with the liposome-triad structure illustrated in Fig. 1 (top) and provide evidence for the mechanism of proton translocation. Excitation of the porphyrin moiety of **1** in liposomes with 5-ns laser pulses at 430 nm (excitation into the intense Soret absorption band for the laser flash experiments was necessary due to the dilute conditions) resulted in formation of C^+-P-Q^- , as detected by the strong transient absorbance of the carotenoid radical cation at 930 nm (refs 13, 14). The yield of this species was ~ 0.1 and its lifetime was ~ 110 ns in the absence of Q_s and ~ 60 ns in liposomes in which Q_s was present in the bilayer. Titration of liposomes containing **1**, but not Q_s , with $Na_2S_2O_4$ (which reduces the triad quinone moiety, thereby blocking photoinduced electron transfer and C^+-P-Q^- formation) decreased the yield of C^+-P-Q^- monotonically to ~ 0 without quenching its lifetime. Because $Na_2S_2O_4$ is not expected to cross the bilayer, these experiments demonstrate that most of the triads must be arranged with their quinone ends proximal to the external aqueous phase. Indirect evidence that the carotenoid moiety extends into the membrane comes from previous studies in which transmembrane photoconduction by triads similar to **1** has been demonstrated^{15,16}. These results imply that the carotenoid spans a substantial portion of the bilayer membrane. Assuming that some fraction of the population of **1** in the liposomes is similarly arranged, the carotenoid must extend inward toward the intraliposomal space, as shown in Fig. 1 (top).

The fluorescence excitation spectrum of the pH-indicating dye PS contained inside the liposomes (see Fig. 1 legend) was measured before and after a 10-min irradiation of a solution of liposomes with a ~ 5 mW cm^{-2} beam of 650-nm light having a 10-nm bandpass. These liposomes contained **1** in the bilayer but did not contain the shuttle quinone Q_s . The two fluorescence excitation spectra are virtually superimposable, indicating that there was no change in the protonation state of the dye as a result of irradiation. Exactly the same results were obtained when liposomes that contained Q_s but no triad **1** were irradiated. Similarly, there was no change in the protonation state of the dye when liposomes containing both **1** and Q_s were kept in the dark or were irradiated at 780 nm (where **1** does not absorb). Finally, liposomes containing **1** and Q_s , but which were treated with $Na_2S_2O_4$ before irradiation, were not photochemically active.

Figure 2 presents excitation spectra of PS in liposomes containing triad **1** and Q_s in the bilayer as a function of irradiation time. The increasing ratio of I_{406}/I_{456} (see Fig. 1 legend) with irradiation time indicates that the dye population is increasingly shifted toward the protonated form. Figure 3 shows the I_{406}/I_{456} ratio for this experiment as a function of irradiation time. In a complementary experiment in which PS was added only to the external aqueous phase, the bulk pH increased with irradiation time. The change in protonation state of an irradiated sample upon addition of 1×10^{-7} mol of a proton ionophore (carbonyl cyanide 4-(trifluoromethoxy)phenylhydrazone (FCCP)) is also shown in Fig. 3. The proton ionophore relaxes the protonation state of the dye to its level before irradiation.

These results demonstrate that in the presence of Q_s , light absorbed by **1** drives protons into the interior aqueous phase of the liposomes. A mechanism for proton translocation consistent with these experiments is presented in the elementary processes 1–7 illustrated in Fig. 1 (top).

Step 1 includes excitation and two-step charge separation. Photoinduced electron transfer and subsequent formation of the carotenoid radical cation is confirmed by the transient absorbance at 930 nm of the C^+-P-Q^- species. The role of C^+-P-Q^- in proton translocation is evidenced by the fact that proton translocation is not observed when the quinone moiety of **1** is reduced before excitation so that C^+-P-Q^- is not formed. In step 2, Q_s near the external aqueous phase accepts an electron from C^+-P-Q^- . This step is thermodynamically allowed (Fig. 1 legend) and is consistent

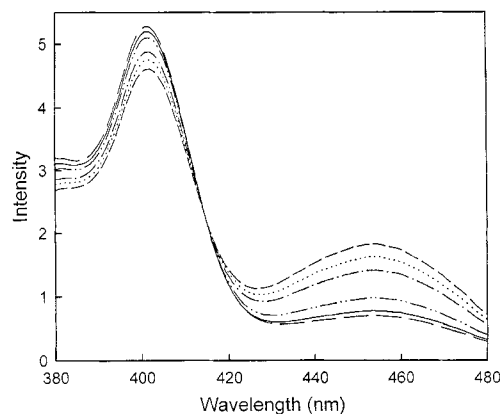


Figure 2 Excitation spectra for emission at 510 nm of PS inside liposomes containing **1** and shuttle quinone Q_s in the bilayer. The liposomes were irradiated for 0 (---), 1 (·····), 3 (— · — · —), 10 (— — — —), 25 (—) and 55 (— — — —) min with ~ 5 mW of 650-nm light. Oxygen and CO_2 were removed from the cuvette by flowing argon.

with the control experiments mentioned above where proton pumping was not observed in the absence of Q_s . In step 3 the Q_s radical anion accepts a proton from the nearby external aqueous phase. This step is driven by the pK of Q_s^- and forms the semiquinone, Q_sH . (Protonation of the anion radical of **1** followed by electron transfer to Q_s and subsequent protonation would also yield Q_sH .) Step 4 indicates the diffusion of Q_sH across the bilayer. Q_sH should be sufficiently nonpolar that it may cross the membrane to the region near the interior aqueous phase where the oxidizing potential (C^+) is located. Step 5 depicts the oxidation of Q_sH by the radical cation of **1**, regenerating the photocatalyst. Evidence for electron donation to the carotenoid radical cation by the reduced quinone is provided by the quenched lifetime of the radical cation (60 versus 110 ns) when Q_s is present in the bilayer. This experiment cannot distinguish between electron delivery by Q_sH or Q_s^- , the latter being part of a futile cycle in that protons are not translocated. In step 6 the protonated quinone deprotonates with a driving force provided by its pK_a (ref. 17) of ~ -6 . Step 7 is included to illustrate the diffusion of Q_s to the exterior region of the bilayer. The sum of steps 1–7 describes photoinduced electron transfer that generates intramembrane redox potential which in turn drives the vectorial translocation of protons by a quinone shuttle.

In order to interpret changes in the excitation spectrum of PS in terms of the quantum yield of proton translocation, we have constructed a model based on three assumptions concerning the response of PS to proton influx. (1) Because the pK of PS (7.2) is well separated from that of the ionizable groups of the phospholipids (~ 2 for the phosphate and serine carboxylate and ~ 9 for the serine amino group, measured by titration of liposomes), the initial protons imported into a liposome will protonate PS. That is, the equilibrium $PS + H_3O^+ \rightleftharpoons HPS + H_2O$ lies far to the right because addition of a single proton to the interior liposome volume would lower the pH to well below 7.2 (in the absence of dye). (2) Before exposure to the actinic light ($t = 0$), the ratio of the number of protonated to unprotonated PS molecules trapped in the interior of a dye-containing liposome, $[(N_{HPS})/(N_{PS})]_{t=0}$, is fixed by the pH of the aqueous phase in which the liposomes were formed (6.8 to 7.2 in different experiments). (3) Because of assumption (1), the response of the dye to the import of n protons is limited to a change in the above ratio to $(N_{HPS} + n)/(N_{PS} - n)$. In effect, due to the small internal volume of the liposomes and the buffer capacity of the dye, the dye is viewed as a proton counter.

The quantum yield of proton translocation was calculated as

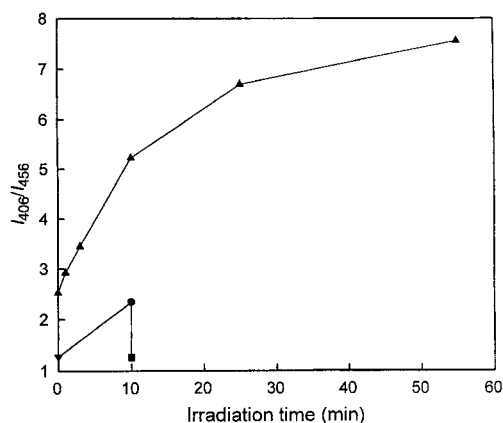


Figure 3 The ratio I_{406}/I_{456} from the data in Fig. 2 plotted as a function of irradiation time (▲). In a different experiment, I_{406}/I_{456} before irradiation (▼), after irradiation, (●) and after addition of 1×10^{-7} mol FCCP (■).

follows. The measured ratio of I_{406}/I_{456} at a given irradiation time was converted into the canonical ratio $(N_{HPS})/(N_{PS})$ by finding the pH corresponding to I_{406}/I_{456} from an experimental titration curve of PS and then solving the equation $pH = pK - \log[(N_{HPS})/(N_{PS})]$. From the ratio $[(N_{HPS})/(N_{PS})]_{t=0}$ and the average number of PS molecules ($N_{HPS} + N_{PS}$) per liposome (~ 50), N_{HPS} and N_{PS} were calculated. After irradiation for time t with the actinic light, the ratio $[I_{406}/I_{456}]_{t=t}$ was measured and converted to $(N_{HPS})/(N_{PS})$ as above. This ratio was set equal to $(N_{HPS} + n)/(N_{PS} - n)$. Solving for n gives the average number of protons translocated per liposome after irradiation for time t . Multiplying n by the total number of liposomes exposed to the actinic light gives the total number of protons translocated. Dividing the total number of protons translocated by the number of quanta absorbed gives an estimate of the quantum yield for proton translocation. After one minute of irradiation, n was calculated to be 1.5, giving a quantum yield of ~ 0.004 .

The assumption that the dye inside the liposome functions as a proton counter leads to a specific prediction: the import of ten protons per liposome results in a calculated change from pH 6.90 to 6.95 in the 1.5 ml external aqueous phase. In an experiment beginning at pH 6.91, 10 min of irradiation resulted in an average of ten protons imported per liposome. In the conjugate experiment, liposomes with PS only in the external phase of pH 6.91 were irradiated for 10 min, giving a new external pH of 6.94. The qualitative agreement of the observed and predicted pH change underpins the view that the dye acts as a proton counter inside the liposomes and demonstrates that proton translocation may be measured by a change in $[H^+]$ in either the internal or external phases.

Because the importation of protons according to the mechanism presented above results in the accumulation of uncompensated charge inside the liposomes, we expect that the importation of multiple protons per liposome would establish a membrane potential which opposes further proton translocation. Typically, a K^+ ionophore such as valinomycin can be used to relax a membrane potential. Indeed, in comparison with the data shown in Fig. 3, after a 10-min irradiation period the I_{406}/I_{456} ratio was 30% larger for liposome suspensions containing 1.7×10^{-5} M valinomycin. This is approximately the increase expected based on a linear extrapolation of the proton translocation after irradiation for 1 min (Fig. 3).

The observation of a photogenerated change in the protonation level of the PS dye inside the liposomes which spontaneously relaxes on addition of a proton ionophore demonstrates that transmem-

brane proton potential was generated and stored in these experiments. We note that in this system photon energy is transduced into vectorial, intramembrane redox potential and then into proton motive force by a chemically cyclic mechanism. That is, it does not require sacrificial electron acceptors or donors and, as in the natural system, the redox potential remains confined to the membrane. Translocation of one proton into a liposome of this type (but lacking PS) generates a calculated proton motive force of > 0.5 pH unit and a concomitant modest membrane potential. It is expected that proton motive force built up across membranes in this way can be harnessed in a variety of biomimetic, energy linked processes including ATP synthesis in liposomes reconstituted with the coupling factor. □

Received 31 May; accepted 25 November 1996.

- Blankenship, R. E., Madigan, M. T. & Bauer, C. E. (eds) *Anoxygenic Photosynthetic Bacteria* (Kluwer Academic, Dordrecht, 1995).
- Gust, D., Moore, T. A. & Moore, A. L. *Acc. Chem. Res.* **26**, 198–205 (1993).
- Wasielowski, M. L. *Chem. Rev.* **92**, 435–461 (1992).
- Gust, D. & Moore, T. A. *Science* **244**, 35–41 (1989).
- Gust, D. *et al. Science* **248**, 199–201 (1990).
- Gust, D. *et al. J. Am. Chem. Soc.* **115**, 11141–11152 (1993).
- Bixon, M. *et al. Isr. J. Chem.* **32**, 449–455 (1992).
- Connolly, J. D. & Bolton, J. R. in *Photoinduced Electron Transfer Part D* (eds Fox, M. A. & Channon, M.) 303–393 (Elsevier, Amsterdam, 1988).
- Maruyama, K., Oksuka, A., Mataga, N. *Pure Appl. Chem.* **66**, 867–872 (1994).
- Rabanal, F., DeGrado, W. F. & Dutton, P. L. *J. Am. Chem. Soc.* **118**, 473–474 (1996).
- Hung, S.-C. thesis, Arizona State Univ. (1995).
- Szoka, F. Jr & Papahadjopoulos, D. *Proc. Natl. Acad. Sci. USA* **75**, 4194–4198 (1978).
- Moore, T. A. *et al. Nature* **307**, 630–632 (1984).
- Land, E. J. *et al. J. Phys. Chem.* **91**, 4831–4835 (1987).
- Seta, P. *et al. Nature* **316**, 653–655 (1985).
- Moore, T. A. *et al. in Photoinduced Charge Separation and Energy Migration in Supramolecular Species* 283–297 (ed. Balzani, V.) (Reidel, Boston, 1987).
- Beauchamp, A. & Benoit, R. R. *Can. J. Chem.* **44**, 1607–1613 (1966).
- Clement, N. R. & Gould, J. M. *Biochemistry* **20**, 1534–1538, 1539–1543 (1981).
- Foster, R. in *Organic Charge-Transfer Complexes* 52 (Academic, London, 1969).
- Petitou, M., Tuy, F. & Rosenfeld, C. *Anal. Biochem.* **91**, 350–353 (1978).

Acknowledgements: We thank K. Sauer for discussions. This work was supported by the US DOE.

Correspondence should be addressed to T.A.M. (e-mail: tmoores@asu.edu).

Low-latitude glacial cooling in the Southern Hemisphere from amino-acid racemization in emu eggshells

Gifford H. Miller*, John W. Magee† & A. J. T. Jull‡

* INSTAAR and Department of Geological Sciences, University of Colorado, Boulder, Colorado 80309-0450, USA

† Archaeology and Natural History, RSPAS, Australian National University, Canberra, ACT 0200, Australia

‡ NSF Radiocarbon Facility, University of Arizona, Tucson, Arizona 85721, USA

The record of natural climate variability over glacial-interglacial timescales provides a framework from which a better mechanistic understanding of the climate system may be derived. But this approach is limited by the number and distribution of well-dated and reliable palaeoenvironmental archives. Particularly vexing is the conflicting evidence for low-latitude cooling during the Last Glacial Maximum, and the apparent synchrony in glacial activity between the hemispheres despite out-of-phase insolation forcing^{1–4}. Here we utilize the temperature-dependent amino acid racemization reaction in radiocarbon-dated emu eggshell fragments from the continental interior of Australia to reconstruct low-altitude subtropical temperatures for the past 45 kyr. Race-

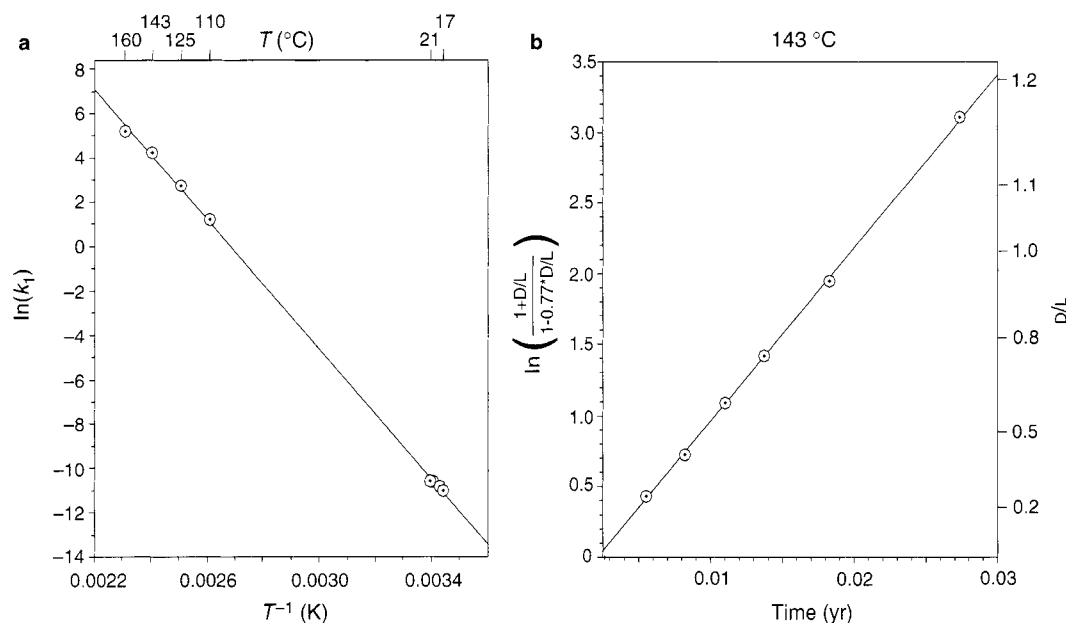


Figure 1 a, Arrhenius plot of isoleucine epimerization in emu eggshell based on simulating time by heating modern emu eggshell isothermally for specific intervals at 160, 143, 125 and 110 °C (for example, Fig. 1b), and calculating an average rate constant for each temperature. Low-temperature points are based on radiocarbon-dated early Holocene samples from a variety of sites across Australia where the mean annual temperature during the historical period is a reasonable approximation (± 1 °C) of the long-term Holocene average. Results of

fit as follows: $\ln(k_1) = 39.33 - 14,653/T$; $r^2 = 0.99$; activation energy $E_a = 29.12$ kcal mol $^{-1}$. **b**, Results of one high-temperature simulation in which modern emu eggshell were heated for specific intervals at 143 °C, illustrating the exceptionally close approximation to reversible first-order kinetics. The forward rate constant, k_1 , was found to be 122.7 yr $^{-1}$ is proportional to the slope of the regression line ($r^2 = 0.99$).

mization rate-changes indicate that millennial-scale average air temperatures were at least 9 °C lower between 45 and 16 kyr BP than since 16 kyr BP. A temperature change of this magnitude, coupled with other low-latitude palaeotemperature records that indicate substantial cooling^{3,4}, must reflect global processes, which, we speculate, involved glacial-age reduction in atmospheric water vapour content.

Australia, situated far from any of the primary ice-age amplifiers of climate change (such as ice–albedo feedback) and lacking major topographic relief, acts as a passive recorder of global-scale climate events. Quaternary environmental change had a major impact on the Australian landscape and biota. The dominant environmental indicators are vegetation change, normally interpreted in terms of changing moisture balance⁵, episodic wet/dry cycles across the semiarid interior^{6,7} and treeline lowering and minor glaciation in Tasmania, in the Snowy Mountains of southeast Australia, and in the mountains of adjacent New Guinea^{8,9}. Insolation changes caused by the Earth's orbital irregularities produce austral summer insolation maxima during Northern Hemisphere glaciations, leading to the speculation that Quaternary temperature change across the interior lowlands was minimal^{8,10}. However, evidence from which past temperatures may be estimated is scanty, especially over the vast continental interior where preservation of typical palaeoclimate proxies, such as pollen, is poor.

One of the most abundant terrestrial fossil materials in the Australian interior is the eggshell of a variety of large birds. Ratite (pertaining to a group of large, flightless birds found on most continents) eggshell, better than any other biomineral, retains indigenous proteinaceous residues over geological time without diffusional loss, or uptake of secondary molecules. Brooks *et al.*¹¹ demonstrated that amino acid racemization in ostrich eggshell follows first-order reversible kinetics nearly to racemic equilibrium, and the method has been used to date archaeological sites beyond the range of radiocarbon dating^{12,13}. Recently we confirmed the

utility of ratite eggshell of the Australian emu (*Dromaius*), and the giant Mhirung, *Genyornis*, an even larger flightless bird that became extinct in Late Pleistocene times. Small sample requirements for both accelerator mass spectrometry (AMS) ^{14}C dating and amino acid analysis allow both analyses on individual fragments weighing as little as 15 mg; typical fossil fragments are 0.1–1 g. The temperature dependence of the racemization reaction is defined by the Arrhenius parameters (Fig. 1a). The close correspondence to reversible first-order linear kinetics under isothermal conditions to a D/L ratio in excess of 1.0 (Fig. 1b; see legend for details of D/L measurements) confirms that the reaction may be used for palaeothermometry or geochronology over a wide range of geological time. Unlike dated faunal and floral assemblages that provide temperature estimates for specific times, chemically based temperature reconstructions reflect the integrated thermal history experienced by each sample¹⁴. If the temporal density of samples is high, unusually complete temperature reconstructions can be derived.

An extensive series of emu eggshell was collected within a 30 km radius of Madigan Gulf, Lake Eyre, the terminal playa in a large interior drainage that occupies more than one-sixth of the Australian continent (Fig. 2). The playa shore is 13 m below sea level; all samples were collected within 30 m vertically of the playa floor. For deeply buried (≥ 2 m depth) samples, the reaction rate reflects mean annual temperature¹⁵. Most of our samples came from the surface of recent deflation hollows, and it is possible that some fragments had been near the surface for several hundred years. Eggshells are susceptible to wind abrasion, and survival at the surface for more than a few decades is unlikely; we restricted our analyses to unabraded fragments. Because of the exponential dependence of reaction rate on temperature, samples exposed to high-amplitude seasonal temperature fluctuations will have a resultant reaction rate equivalent to a higher temperature than for deeply buried samples. We have modelled the effect on racemization rate of 500 years exposure to high-amplitude diurnal and annual heating using daily

temperature measurements at several depths in desert regions; this uncertainty is incorporated by using asymmetrical error bars in Fig. 2.

In Fig. 2a we plot calibrated age against corresponding D/L ratio measured in 42 individual fragments of emu eggshell and 3 of *Genyornis* from the Madigan Gulf region. As plotted, the data should fall on a straight line if temperature is constant. The prominent kink in the data indicate a substantial temperature shift occurred about 16,000 calendar years before present (16 cal.kyr BP). To quantify this change, we fit two sets of regression lines to the data. The thin line represents a linear regression from the modern samples to ~16 cal.kyr ago and a second linear regression through the remainder of the data, both assuming random errors. An alternative and more conservative strategy is to assume that most of the Holocene samples have been shallowly buried and only the lowest D/L ratios should be used to define the regression line; the steep thick line in Fig. 2a is based on a regression through modern samples and five latest Pleistocene samples that were collected from deeply buried contexts. For the older portion of the curve, we have eliminated all Last Glacial Maximum (LGM) samples (17–28 kyr) from consideration. This was a time of lake-floor deflation in the Madigan Gulf region, and most LGM samples come from a thin, discontinuous playa-marginal unit formed by deflation of playa sediments during an interval of hyper-aridity¹⁶. Other LGM aeolian units are also thin; consequently all LGM samples are presumed to have been always near the surface. In contrast, the samples dated between 28 and 47 cal.kyr BP were collected from recent deflation hollows in thick aeolian sequences, hence are more likely to have been deeply buried. The two regression lines intersect at about 17 cal. kyr BP.

McCoy¹⁷ has previously shown that the most robust quantitative estimate from D/L ratios is the temperature-difference calculation, with a precision of $\pm 1^\circ\text{C}$. The difference in temperature manifest by the slope changes in Fig. 2a can be calculated from the Arrhenius parameters in Fig. 1a. Assuming random errors (thin lines) the difference in slopes corresponds to a 20 °C Pleistocene temperature reduction. We argue that this figure is inflated by the shallow burial of LGM samples compared with deeply buried older eggshell, and a systematic bias in the Holocene series owing to shallow burial histories. Assuming systematic errors (thick lines), the temperature difference decreases to 9 °C, a value that, we argue, better reflects the uncertainties caused by variable burial histories. In either case, the data indicate a dramatic cooling of the continental interior of Australia during the last glaciation.

To test this conclusion, we obtained a second dataset from eggshell buried in a dune system fringing Lake Victoria, where the current mean annual temperature is 4 °C lower than the Madigan Gulf region (Fig. 2b). Two collections dated to ~16 cal. kyr BP were obtained from deeply buried sites, hence have small uncertainties. A linear fit between modern and 16 kyr corresponds to a temperature of 18 °C, similar to the current mean annual temperature (17.3 °C; ref. 18). A best-fit line from 16 kyr to the older samples has a dramatically lower slope, equivalent to a temperature lowering of 10 °C. Although the D/L ratios are significantly lower in the cooler Lake Victoria collections than equivalent-aged samples from the Madigan Gulf region, the implied temperature depression for the two sites is the same, supporting our initial conclusion, and suggesting that kinetic irregularities are unlikely to be a contributing factor. Note, for example, that the entire Lake Victoria data set has D/L ratios that fall within the initial linear results at Lake Eyre.

A similar data set is available from seven paired ¹⁴C-D/L measurements on emu eggshell from the Lake Mungo region, situated 150 km northeast of Lake Victoria (these data are not shown, but are available as Supplementary Information). Because the burial history of these samples is uncertain, a rigorous evaluation of the thermal history is not possible. However, regression lines through the data, assuming one anomalous sample was shallowly buried, indicates a Pleistocene temperature reduction of 8–9 °C.

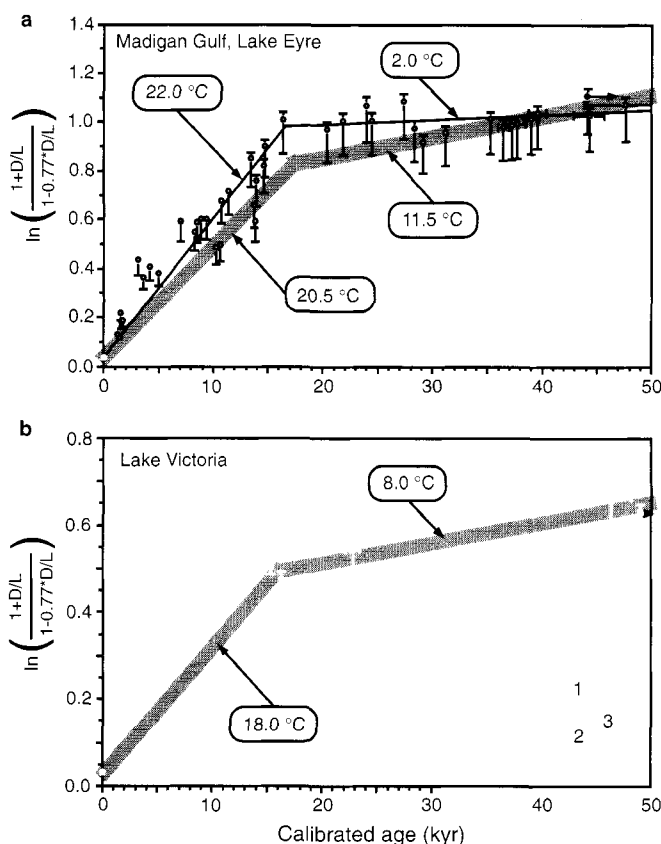


Figure 2 Extent of isoleucine epimerization plotted against corresponding calibrated age for 45 eggshell fragments from Lake Eyre (a) and 6 samples from Lake Victoria (b). Asymmetric vertical error bars incorporate the possibility of shallow burial. Horizontal error bars reflect $\pm 1\sigma$ dating errors where the uncertainty exceeds the diameter of the plotted point. Closed circles, data from emu eggshells; open circles, *Genyornis* eggshell D/L ratios converted to equivalent emu eggshell ratios by multiplying by 1.18, the average ratio of the two taxa from sites where they are found together. Derivation of regression lines is discussed in the text. The data on which these graphs are based are available as Supplementary Information. Inset shows the Lake Eyre Basin (hatch region), Madigan Gulf (1), Lake Victoria (2) and Lake Mungo (3).

Previous estimates for LGM temperature reductions in terrestrial Australasia are based on evidence restricted to mountainous regions, principally the magnitude of snowline and treeline depressions, and the occurrence of diagnostic periglacial phenomena. LGM snowline depressions are estimated to have been ~1,000 m in Tasmania and New Guinea and 850 m in southeast Australia. Assuming no change in lapse rate, a glacial maximum temperature depression of ~6 °C has been estimated^{8,9}, although the actual depression could have been substantially greater if precipitation was less. Treeline depression in the same regions ranges from 1,200 to 1,700 m; estimates of corresponding temperature depressions range from 6 to 10 °C (refs 19, 20). The occurrence of scree slopes, blockfields and permafrost in the mountains of New South Wales suggests a LGM temperature depression of 7–11 °C (ref. 21). These reconstructions for the mountainous regions conflict strongly with CLIMAP¹ and more recent^{22,23} sea surface temperature (SST) reconstructions for the surrounding seas at the LGM which average <2 °C below modern around northern Australia, a discrepancy widely noted^{3,24–26}. The lack of secure LGM palaeotemperature estimates for the low-elevation interior of Australia has left this discrepancy unresolved.

Our data provide quantitative LGM temperature estimates for the

vast arid and semiarid Australian interior. Because the chemical temperature reconstructions from the racemization evidence are annual integrations rather than seasonal averages, they place severe constraints on climate reconstructions, unlike vegetation- and snowline-based estimates that record specific events or short-term seasonal averages. Additionally, inferences for sea-level temperatures drawn from high-elevation proxy data are limited owing to possible changes in lapse rates at the LGM, an uncertainty avoided by our data set, which was acquired from sea level near the centre of the continent and is independent of lapse rate variations.

The racemization rate changes observed in emu eggshell are, to our knowledge, the first indications that dramatic Pleistocene cooling occurred across Australia, and that this cooling began well before the LGM. Because the reaction rate exhibits an exponential dependency on temperature, brief warm intervals would disproportionately influence the calculated average temperature reduction. Consequently, we can conclude that the entire interval from at least 45 cal.kyr to 16 cal.kyr BP was characterized by temperatures consistently at least 6 °C lower than present, allowing that the LGM may have been substantially colder than the average 9 °C reduction. Our results also demonstrate that the onset of warming in Australia, and by inference globally, occurred well before the start of the Holocene. We estimate that rapid warming, to near modern values, occurred 16 ± 1 cal.kyr BP, coincident with vegetation change in southern South America²⁷, major changes in ocean circulation and Northern Hemisphere ice-sheet recession. These inferences are consistent with recent evidence from low-elevation, low-latitude sites elsewhere, including Barbados corals³ and changes in noble gases in Pleistocene ground water from Brazil, southern Australia and Namibia⁴, and racemization in Jamaican landsnails²⁸, that also indicates large LGM temperature reductions.

Although the primary forcing of Quaternary climate change is widely accepted to result from systematic changes in the distribution of solar insolation caused by irregularities in Earth's orbital parameters, these relatively modest changes in energy are amplified by a variety of powerful feedback mechanisms associated with ice-sheet growth and decay. The Australian continent lies far from the regions of strong local ice/snow feedback, and Madigan Gulf was more than 750 km from the Pleistocene coastline, minimizing any direct oceanic cooling. The dramatic temperature reduction implied by our data, coupled with other low-latitude proxies that also indicate substantial subtropical cooling^{3,4}, require the cooling to be a global phenomenon.

When forced by the CLIMAP SST reconstructions¹, climate models have been unable to simulate terrestrial cooling as large as that we have reconstructed. However, the minor lowering of tropical SSTs estimated by CLIMAP has been recently challenged by Guilderson *et al.*³. This suggests that it may be imprudent to rely on climate models forced by CLIMAP SSTs when attempting to identify the mechanisms of glacial cooling. Only a few climate model simulations of glacial climate have been conducted in which SSTs are predicted by the model. One such recent LGM simulation by Webb *et al.*²⁹, using modern values for ocean heat transport, yields temperatures 9 °C lower than present over the Australian interior. These results are in closer agreement to our reconstructed temperatures than are the earlier simulations with atmospheric general circulation models based on CLIMAP SSTs. They indicate that a glacial ocean with ocean heat transport similar to today, possibly associated with a vigorous North Atlantic Intermediate Water circulation³⁰, would produce substantially lower temperatures in the tropics and Southern Hemisphere when forced by radiative effects of increased terrestrial ice, decrease in ocean area, and reduction of atmospheric CO₂ concentration. We speculate that tropical SSTs substantially lower than those indicated by the CLIMAP reconstructions might result in a reduction in atmospheric water vapour content that, combined with other

boundary conditions changes, effectively coupled the two hemispheres and amplified global cooling. Our data indicate that the combination of circumstances that result in substantial global cooling occurred not only during the LGM, but was in fact initiated much earlier, during marine isotope stage 3, and persisted until about 16 cal.kyr BP. □

Methods

Before analysis, each sample was mechanically ground to remove surface contaminants. For ¹⁴C dating, grinding was followed by the removal of 50% of the remaining mass by the stoichiometric addition of 2M HCl. For samples expected to be >30 kyr old, we removed 90% of the original mass and stored the cleaned samples under inert gas to minimize contamination by diffusion of younger carbon into the calcite crystals. Emu eggshell collected live in our study region in 1991 AD had a ¹⁴C activity consistent with the 1991 atmosphere; eggshell collected from a nest in New South Wales in 1897 AD gave a radiocarbon age of 105 ± 75 yr BP. These results indicate that no 'old' carbon is incorporated during calcification of the eggshell. All radiocarbon dates have been converted to calendric ages following Stuiver and Reimer³¹ for samples <10 kyr BP, and a polynomial fit of U-Th versus ¹⁴C ages (E. Bard, unpublished data) for samples >10 kyr BP. The extent of racemization was quantified using high-pressure liquid chromatography. After mechanical cleaning, an additional 33% of the remaining mass was removed by 2M HCl. A small subset of the cleaned fragment was decalcified and hydrolysed in excess 6M HCl under nitrogen for 22 h at 110 °C. L-isoleucine and its epimerization product, D-alloisoleucine, were separated by ion-exchange chromatography, and quantified by post-column derivitization with O-phthalaldehyde and ultraviolet detection. L-isoleucine/D-alloisoleucine (or D/L) ratios were calculated as the ratio of peak heights after electronic integration. Replicate analyses were performed on all samples; typical analytical uncertainty was $\pm 1\%$. Modern emu eggshell have a D/L ratio of about 0.02, and at racemic equilibrium the ratio is 1.30.

Received 7 February; accepted 6 November 1996.

1. CLIMAP Project Members *Map Chart Series MC-36* (Geol. Soc. America, Boulder, 1981).
2. Berger, A. & Loutre, M. F. *Quat. Sci. Rev.* **10**, 297–317 (1991).
3. Guilderson, T. P., Fairbanks, R. G. & Rubenstein, J. L. *Science* **263**, 663–665 (1994).
4. Stute, M. *et al. Science* **269**, 379–383 (1995).
5. Kershaw, A. P. *et al. A Comparison of Long Quaternary Pollen Records from the Atherton and Western Plains Volcanic Provinces, Australia* (Geol. Soc. Australia, Sydney, 1991).
6. Magee, J. W., Bowler, J. M., Miller, G. H. & Williams, D. L. G. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **113**, 3–42 (1995).
7. Nanson, G. C., Chen, X. Y. & Price, D. M. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **118**, 87–102 (1995).
8. Colhoun, E. A. & Fitzsimons, S. J. *Quat. Sci. Rev.* **9**, 199–216 (1990).
9. Hope, G. S. & Peterson, J. A. *Bull. R. Soc. NZ* **13**, 155–162 (1975).
10. Bowler, J. M. in *Quaternary Studies* (eds Suggate, R. P. & Cresswell, M. M.) 75–82 (Royal Soc. New Zealand, Wellington, 1975).
11. Brooks, A. *et al. Science* **248**, 60–64 (1990).
12. Henry, D. O. & Miller, G. H. *Paleorient* **18/2**, 45–52 (1992).
13. Brooks, A. S., Hare, P. E. & Kokis, J. E. *Y Carnegie Inst. Wash.* **92**, 80–85 (1993).
14. Miller, G. H. *et al. Nature* **326**, 593–595 (1987).
15. Miller, G. H. & Brigham-Grette, J. *Quat. Int.* **1**, 111–128 (1989).
16. Magee, J. W. & Miller, G. H. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* (in the press).
17. McCoy, W. D. *Quat. Sci. Rev.* **6**, 43–54 (1987).
18. Hutchinson, M. F. in *Data Assimilation Systems* (ed. Jasper, J. D.) 104–113 (Bureau of Meteorol., Melbourne, 1991).
19. Hope, G. S. in *CLIMANZ 3, Proc. 3rd Symp. on the Late Quaternary Climatic History of Australasia* (eds Donnelly, T. H. & Wasson, R. J.) 91–99 (CSIRO Div. of Water Resources, Melbourne Univ., 1989).
20. Singh, G. & Geissler, E. A. *Phil. Trans. R. Soc. Lond. B* **311**, 379–447 (1985).
21. Galloway, R. W. *J. Geol.* **73**, 603–618 (1965).
22. Anderson, D. M., Prell, W. L. & Barratt, N. J. *Paleoceanography* **4**, 615–627 (1989).
23. Thunell, R., Anderson, D., Gellar, D. & Miao, Q. *Quat. Res.* **41**, 255–264 (1994).
24. Rind, D. & Peteet, D. *Quat. Res.* **24**, 1–22 (1985).
25. Webster, P. & Streten, N. *Quat. Res.* **10**, 279–309 (1978).
26. Smith, I. N. in *CLIMANZ 3, Proc. 3rd Symp. Late Quaternary Climatic History of Australasia* (eds Donnelly, T. H. & Wasson, R. J.) 33–41 (CSIRO, Melbourne, 1989).
27. Markgraf, V. *et al. Clim. Dynam.* **6**, 193–211 (1992).
28. Goodfriend, G. A. & Mitterer, R. M. *Geol. Soc. Am. Bull.* **105**, 18–29 (1993).
29. Webb, R. S., Lehman, S. J., Rind, D., Healy, R. J. & Sigman, D. *Nature* (submitted).
30. Sigman, D. M. & Lehman, S. J. (suppl.) *Eos* **76**, F285 (1995).
31. Stuiver, M. & Reimer, P. *Radiocarbon* **35**, 215–230 (1993).

Supplementary Information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from Mary Sheehan at the London editorial office of Nature.

Acknowledgements: We thank C. Hart, B. Roark, A. Rosewater and L. Toolin for assistance in the laboratory; G. Atkin, P. Clark, P. DeDecker, C. Hart, B. Johnson, O. Miller and S. Webb for assistance in the field, and the Mitchell families for access to Muloorina Station. This project was supported by the US NSF (G.H.M., A.J.T.) and the Australian National University (J.W.M., G.H.M.).

Correspondence should be addressed to G.H.M. (e-mail: gmler@colorado.edu).

Seismic structure of the Iceland mantle plume

Cecily J. Wolfe*, Ingi Th. Bjarnason†, John C. VanDecar* & Sean C. Solomon*

* Department of Terrestrial Magnetism, Carnegie Institution of Washington, 5241 Broad Branch Road, NW, Washington DC 20015, USA

† Science Institute, University of Iceland, Reykjavik, Iceland

Oceanic hotspots are generally accepted to be the manifestations of plumes of hot, upwelling mantle material^{1,2}, but the nature of such flows remains enigmatic. Iceland, for example, is one of the most thoroughly investigated hotspots, yet previous seismological^{3–5} and geodynamic^{6–12} studies have been unable to constrain the width or temperature of the plume. Here we report the results of a regional broadband seismic experiment undertaken to determine the three-dimensional velocity structure of the upper mantle beneath Iceland using relative travel times of body waves from teleseismic earthquakes. Inversion solutions of the data show a cylindrical zone of low P- and S-wave velocities that extends from 100 km to at least 400 km depth beneath central Iceland. The radius of the low-velocity anomaly is about 150 km, and its magnitude is approximately 2% for P waves and 4% for S waves, indicating that Iceland is underlain by a hot, narrow plume of upwelling mantle.

Although the interaction between the Iceland plume and the Mid-Atlantic Ridge (MAR) has several evident manifestations, a detailed understanding of that interaction has been elusive. The Iceland hotspot generates a broad topographic high⁶, thicker than normal crust¹³, and a pronounced geochemical⁷ anomaly that extends along the adjacent Reykjanes ridge. Neovolcanic zones within Iceland, interpreted as on-land extensions of the MAR, have similarities to oceanic spreading centre segments (Fig. 1). On the basis of the excess crustal thickness^{8,9}, topographic anomaly^{9,10} and rare-earth element patterns⁹, kinematic models for mantle flow beneath Iceland and the Reykjanes ridge suggest a narrow plume radius (100 km) and a significantly hotter temperature ($\Delta T = 200^\circ\text{C}$) than the surrounding mantle. In contrast, recent three-dimensional variable-viscosity models of the dynamics of mantle flow and melting beneath this ridge-centred hotspot demonstrate that a much broader (300 km radius) and cooler ($\Delta T = 75^\circ\text{C}$) plume can match the observed topographic and gravity anomalies^{11,12}. Whether the shallow horizontal component of mantle flow is radially symmetric about the hotspot centre^{11,12,14} or is channelled along the Reykjanes ridge axis^{6,9,10} has not been ascertained. Global seismological models^{3,4} display lower than average shear velocities in the upper mantle surrounding Iceland, but the lateral resolution of such models is no better than $\sim 1,000$ km. Tryggvason *et al.*⁵ inverted arrival times of teleseismic P waves recorded on an unevenly distributed network of short-period analogue seismic stations and found a narrow low-velocity anomaly in the upper 400 km beneath central Iceland, but the structure and magnitude of the anomaly were poorly constrained.

ICEMELT¹⁵ is a regional broadband seismic experiment, initially undertaken by the Carnegie Institution of Washington and now run in cooperation with the Science Institute of the University of Iceland, designed to determine the detailed upper-mantle structure beneath a mid-ocean-ridge hotspot. The ICEMELT network, installed during 1993–95 and in operation until autumn 1996, consists of 15 portable broadband, three-component STS-2 seismometers distributed throughout Iceland (Fig. 1). Relative delay times of filtered P and S waves from teleseismic earthquakes have been measured¹⁵ with a multichannel cross-correlation technique¹⁶. The resulting set of delay times is of high accuracy (standard errors

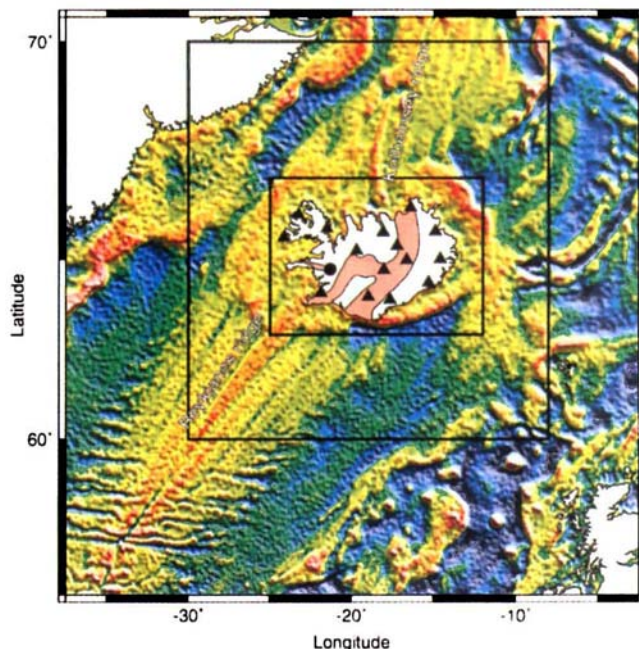


Figure 1 Location of ICEMELT seismic stations (filled triangles) within Iceland. We have also used data from the Global Seismographic Network station BORG in western Iceland (filled circle). The regional setting is illustrated by the satellite-derived gravity field²⁹ of the North Atlantic; yellow-to-red colours indicate more positive gravity values, while green-to-blue colours indicate more negative values. Positive gravity anomalies generally coincide with shallower than normal sea floor. Icelandic neovolcanic zones are shaded in red. The outer and inner rectangles represent respectively the exterior boundary, in map view, of the seismic velocity grid used for delay-time inversion and the interior boundary where the best resolution is obtained and within which we confine our interpretations. The model grid extends from 0 to 1,000 km in depth.

of ~ 30 ms for P waves and ~ 100 ms for S waves). As described by Bjarnason *et al.*¹⁵, there is a consistent qualitative pattern for measured body-wave delays, which indicates a narrow low-velocity anomaly throughout the upper mantle beneath central Iceland: P and S waves that travel through this region are significantly delayed relative to waves traversing other paths.

Here P- and S-wave delays are independently used to solve for three-dimensional velocity structure, earthquake relocations and station terms by means of a robust nonlinear, regularized travel-time inversion scheme¹⁷. Because these types of problems are inherently underdetermined, this method searches for the minimum-structure model required to satisfy the data by minimizing spatial gradients and roughness. The inversions incorporate relative arrival times of 601 P waves from 86 earthquakes (including 12 earthquakes with core phases) and 560 S-wave times from 78 earthquakes (including 13 earthquakes with core phases). The earthquakes employed in this study are well distributed in azimuth around Iceland (see Supplementary Information). The dominant seismic wavelengths are ~ 10 km for the P waves and 50–75 km for the S waves.

The inversion solutions, with one nonlinear iteration to incorporate ray bending effects, are shown in Fig. 2 as percentage velocity perturbations to a standard radial Earth model¹⁸. Our models explain 95% of the root-mean-squared (r.m.s.) residual for P-wave delay data (from 0.4 to 0.02 s) and 91% of that for S-wave data (from 1.4 to 0.05 s), with final r.m.s. values slightly smaller than our *a priori* estimates of data errors. Because the station spacing is ~ 75 km and incidence angles of teleseismic waves steepen in the shallowest mantle, there are no crossing wave paths above ~ 100 km depth¹⁷. Velocity structure is thus poorly constrained at these shallow depths, although the station terms shown in Fig. 2 account

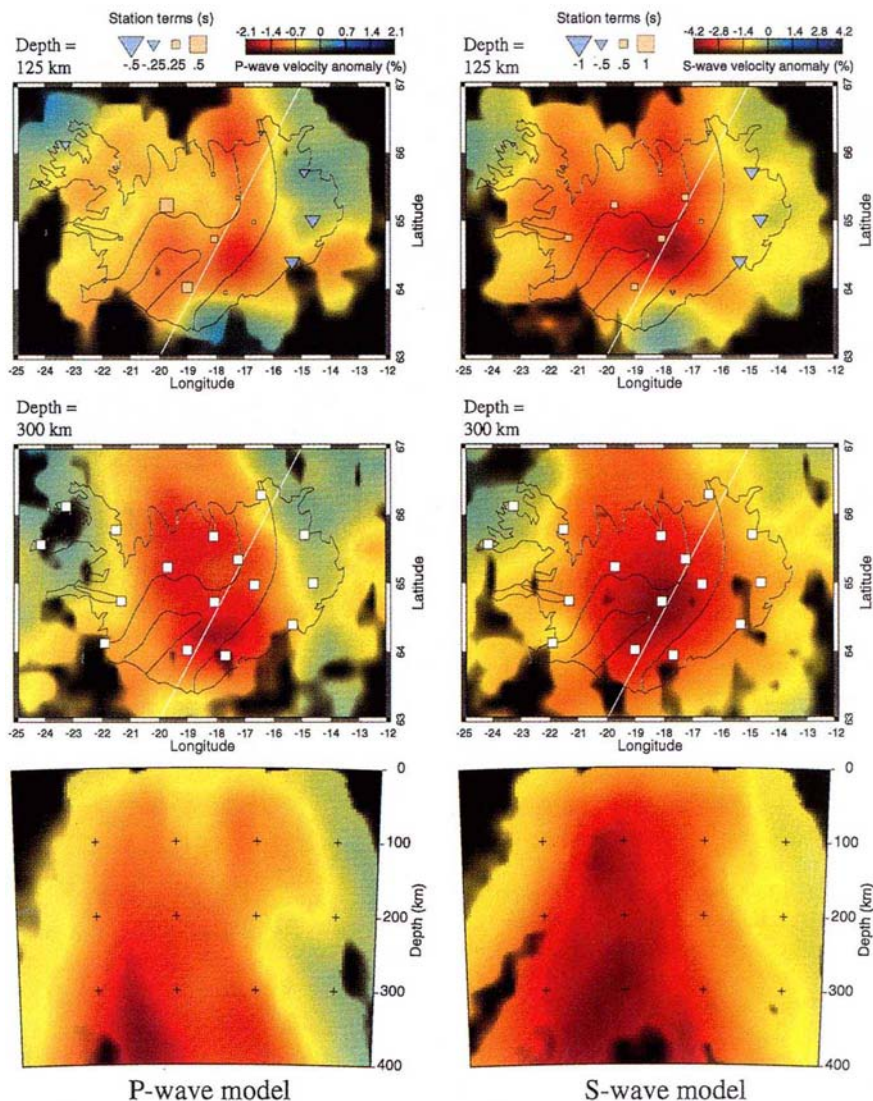


Figure 2 Cross-sections through the P-wave (left-hand column) and S-wave (right-hand column) models. Within this region, velocity nodes are spaced at 25 km in depth, 0.5° in longitude and 0.25° in latitude. Velocity perturbation models are shown in map view at depths of 125 km (top row) and 300 km (middle row); vertical cross-sections (bottom row) are along the white line in map views. The velocity perturbations shown are relative variations across the modelled volume; the absolute velocities are unconstrained. Yellow-to-red colours represent low-velocity anomalies: regions with sparse coverage are faded to black. Station locations are shown as white squares in the middle row. Square and triangular

symbols in the top row indicate the magnitude and sign of the station terms. The inversions solve simultaneously for independent (undamped) station terms to absorb as much shallow structure as possible into these terms and to avoid projecting such structure into the underlying velocity model. These station-term times represent the vertical travel-time of waves associated with each station and thus account in part for the integrated differences in mantle structure above 100 km depth, combined with the effects of crustal thickness variations and differences in station elevation. The Icelandic coastline and neovolcanic zone boundaries are shown as black lines.

in part for the integrated differences in mantle structure above 100 km depth, combined with the effects of crustal thickness variations and differences in station elevation.

The most prominent feature in our inversion solutions is a low-velocity anomaly beneath central Iceland extending down to at least 400 km depth (Fig. 2). At 300 km depth, the low-velocity anomaly is approximately circular and 150–200 km in radius (defined by the region at which the perturbation drops below $1/e$ of the maximum value). At 125 km depth, in contrast, the low-velocity region coincides more nearly with the surface expression of the neovolcanic zone, the correlation being most evident in northern Iceland. The centre of the low-velocity region is vertically continuous and does not show significant lateral shifts over the depth range 100–400 km (Fig. 2), in contrast to previous P-wave images of Tryggvason *et al.*⁵ that suggested a more variable structure. We note, however, that these previous results achieved only an 18% reduction

in r.m.s. delays (from 0.46 to 0.38 s)⁵, suggesting that larger errors for the travel times obtained from analogue recordings may have influenced the model.

We have performed a number of resolution tests (see Supplementary Information), which indicate that the basic geometry of both the circular low-velocity anomaly throughout the depth range 100–400 km and the shallower (<150 km) low-velocity anomaly extending along the neovolcanic zones are both well resolved. These tests demonstrate that the inversion solutions (Fig. 2) cannot be produced by either a significantly shallower (≤ 200 km depth) or a broader (300 km radius) low-velocity anomaly. Inversions in which the depth extent of the velocity anomaly is constrained to be less than 400 km do not satisfy the data as well as models that allow deeper structure. In general, the resolution tests show that a narrow, cylindrical slow region extending throughout the Icelandic upper mantle will appear broadened (by as much as 50%) in the inversion

solutions, but there should be no significant artificial widening with depth. These tests also indicate that the magnitudes of the velocity anomalies in the inversion solutions are probably underestimated (we retrieve 75% of the velocity anomaly for the central plume structure). At seismic frequencies, the anomaly magnitude could be further reduced by wavefront healing¹⁹, where rays diffract around the edge of a low-velocity anomaly and arrive before the Fermat ray, reducing the observed time delay by an amount that depends on several factors, including the seismic wavelength, the distances of the wave path from the velocity anomaly and the anomaly magnitude²⁰. Taking into consideration all of these effects, the imaged velocity anomalies (Fig. 2) should be regarded as providing minimum values for the magnitude of the velocity perturbations beneath Iceland and a maximum value for the width of the central low-velocity anomaly.

Our P- and S-wave models are generally similar in structure (Fig. 2), but the magnitude of the anomaly is substantially larger for the S-wave model (4%) than for the P-wave model (2%). We have confirmed this S-to-P anomaly ratio in the raw delay-time data. The relative delays of P and S waves at pairs of stations from common earthquake sources are correlated, with the S delays generally four times larger than the P delays. This ratio is approximately a factor of two larger than if the fractional variations in velocity were equal for P and S waves. There is some difference between the P and S velocity structures beneath central Iceland, which is the point of minimum velocity for S waves, but not for P waves (Fig. 2). As shear-wave splitting of the order of 1–2 s has been observed at several ICEMELT stations¹⁵, mantle anisotropy could account for some of the differences in P- and S-wave models. In particular, a narrow zone of upwelling flow beneath central Iceland would tend to induce vertical alignment of the crystallographic *a*-axes of mantle olivine crystals, which would produce a vertically oriented fast direction of anisotropy for P waves but not for S waves²¹. Our P- and S-wave models also indicate some asymmetry in the integrated structure above 100 km depth, as shown in both the station terms and shallowest mantle structure (Fig. 2), suggesting that the integrated travel times for shallow structure beneath stations on older crust (>7 Myr; ref. 22) along the eastern coast are fast relative to stations on older crust in western Iceland. When combined with the recent observation of the relatively large 35-km crustal thickness in eastern Iceland²³, these results suggest that beneath eastern Iceland the mantle shallower than 100 km displays the highest average wave speeds in the region.

The low-velocity anomaly in the upper mantle beneath Iceland can be interpreted as the locus of the active plume and its interaction with the spreading plate boundary. (Note that we cannot constrain the structure outside of the imaged region, so the full areal extent of mantle affected by the Iceland hotspot remains unknown.) The imaged location of the centre of the plume is broadly similar to the zone of highest ³He/⁴He concentrations within the neovolcanic zones²⁴, which has previously been suggested to mark the region of greatest plume influence. The seismic images indicate that the Iceland mantle plume persists down to at least 400 km depth. Although our data cannot resolve the signature of the plume at greater depth, observations of P-to-S conversions from upper-mantle discontinuities beneath Iceland are consistent with elevated temperatures and upwelling flow in the depth interval 400–700 km (ref. 25).

The maximum velocity anomalies can be converted to a contrast in temperature ΔT between the plume and surrounding mantle, on the basis of laboratory measurements of the temperature derivatives of wave speeds in olivine²⁶. The P- and S-wave anomalies yield ΔT values of 300 and 600 °C, respectively. The significantly larger magnitude of ΔT estimated from the S-wave model could reflect several influences not taken into account in this simple conversion, including the presence of melt²⁷, anelasticity²⁸ and anisotropy²¹. Accounting for the effect of anelastic dispersion²⁸, for instance, gives lower ΔT values of 200 °C for P waves and 300 °C for S waves, results more consistent with previous estimates for Iceland^{8–10,12,13}. Further

analysis of other seismic phases (for example, surface waves, shear-wave splitting) will be necessary to determine the relative importance of anisotropy, anelasticity and melt on the velocity structure.

Both the P- and S-wave images nonetheless provide strong support for the model^{8–10} of a hot and narrow plume beneath central Iceland having a vertical extent of at least 400 km, and against the model of a relatively cool, broad plume, predicted by several recent numerical models^{11,12}. These numerical studies however, indicate that a narrow plume significantly overpredicts the crustal thickness beneath Iceland and the elevation contrast between Iceland and the adjacent mid-ocean ridge. A possible means for reconciling the seismic observation of a narrow Icelandic plume with the results of flow models^{11,12} is the horizontal channelling of plume-derived melts down the nearby ridge^{6,12}. □

Received 1 July; accepted 4 December 1996.

1. Wilson, J. T. *Sci. Am.* **208**, 86–100 (1963).
2. Morgan, W. J. *Nature* **230**, 42–43 (1971).
3. Zhang, Y.-S. & Tanimoto, T. *J. Geophys. Res.* **98**, 9793–9823 (1993).
4. Grand, S. P. *J. Geophys. Res.* **99**, 11591–11621 (1994).
5. Tryggvason, K., Husebye, E. S. & Stefánsson, R. *Tectonophysics* **100**, 94–118 (1983).
6. Vogt, P. R. *Nature* **240**, 338–342 (1972).
7. Schilling, J.-G. *Nature* **242**, 565–571 (1973).
8. McKenzie, D. J. *Petrol.* **25**, 713–765 (1984).
9. White, R. S., Bown, J. W. & Smallwood, J. R. *J. Geol. Soc. Lond.* **152**, 1039–1045 (1995).
10. Schilling, J.-G. *Nature* **352**, 397–403 (1991).
11. Ribe, N. M., Christensen, U. R. & Theissing, J. *Earth Planet. Sci. Lett.* **134**, 155–168 (1995).
12. Ito, G., Lin, J. & Gable, C. W. *Earth Planet. Sci. Lett.* **144**, 53–74 (1996).
13. Bjarnason, I. Th., Menke, W., Flóvenz, Ó. G. & Caress, D. J. *Geophys. Res.* **98**, 6607–6622 (1993).
14. Feighner, M. A., Kellogg, L. H. & Travis, B. J. *Geophys. Res. Lett.* **22**, 715–718 (1995).
15. Bjarnason, I. Th., Wolfe, C. J., Solomon, S. C. & Gudmundson, G. *Geophys. Res. Lett.* **23**, 459–462 (1996); *Geophys. Res. Lett.* **23**, 903 (1996).
16. VanDecar, J. C. & Crosson, R. S. *Bull. Seismol. Soc. Am.* **80**, 150–169 (1990).
17. VanDecar, J. C., James, D. E. & Assumpção, M. *Nature* **378**, 25–31 (1995).
18. Kennett, B. L. N. & Engdahl, E. R. *Geophys. J. Int.* **105**, 429–465 (1991).
19. Wielandt, E. in *Seismic Tomography* (ed. Nolet, G.) 85–98 (Reidel, Dordrecht, 1987).
20. Nataf, H.-C. & VanDecar, J. *Nature* **364**, 115–120 (1993).
21. Kendall, J.-M. *Geophys. Res. Lett.* **21**, 301–304 (1994).
22. Saemundsson, S. *Jökull* **29**, 7–28 (1979).
23. White, R. S. *et al. Eos* **77**, 197, 200–201 (1996).
24. Kurz, M. D., Meyer, P. S. & Sigurdsson, H. *Earth Planet. Sci. Lett.* **74**, 291–305 (1985).
25. Shen, Y., Solomon, S. C., Bjarnason, I. Th. & Purdy, G. M. *Geophys. Res. Lett.* **23**, 3527–3530 (1996).
26. Isaak, D. G. *J. Geophys. Res.* **97**, 1871–1885 (1992).
27. Faul, U. H., Toomey, D. R. & Waff, H. S. *Geophys. Res. Lett.* **21**, 29–32 (1994).
28. Karato, S. *Geophys. Res. Lett.* **20**, 1623–1626 (1993).
29. Smith, W. H. & Sandwell, D. T. *Eos* **76**, 156 (1995).

Supplementary Information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from Mary Sheehan at the London editorial office of Nature.

Acknowledgements: We thank Björn Bjarnason, Birgir Bjarnason, Bryndís Brandsdóttir, Haukur Brynjólfsson, Kristinn Egilsson, Gunnar Gudmundson, Eythór Hannesson, Tryggvi Hardarson, Lárus Helgason, Bogi Ingimundarson, Haraldur Jónsson, Einar Kjartansson, A. Kuehnelt, R. Kuehnelt, Pálmi Sigurdsson, Ragnar Thordmarsson, and the staff of the National Electric Company of Iceland (Landsvirkjun) for assistance with station installation and maintenance; the IRIS-DMS for making available data from the Global Seismograph Network station BORG; W. Menke for discussions; and N. Ribe for comments. This work was supported by the US NSF.

Correspondence should be addressed to C.J.W. (e-mail: cecily@dtm.ciw.edu).

Nest and egg clutches of the dinosaur *Troodon formosus* and the evolution of avian reproductive traits

David J. Varricchio^{*}, Frankie Jackson[†], John J. Borkowski[‡] & John R. Horner[†]

^{*} Old Trail Museum, PO Box 919, Choteau, Montana 59422, USA

[†] Museum of the Rockies, Montana State University, Bozeman, Montana 59717, USA

[‡] Department of Mathematical Sciences, Montana State University, Bozeman, Montana 59717, USA

Living archosaurs (crocodilians and birds) share several reproductive features, including hard-shelled eggs¹, parental care^{2,3}, assembly-line oviducts⁴ and luteal morphology⁵. Nevertheless, crocodilians produce many small eggs that they ovulate, shell

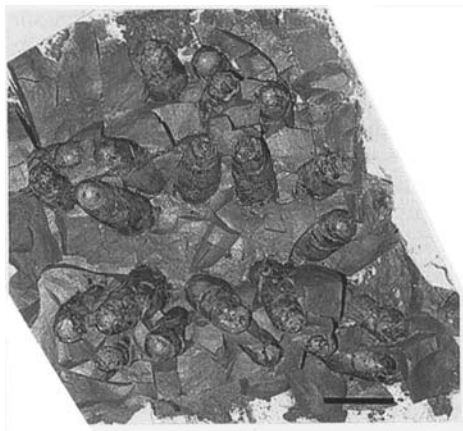


Figure 1 MOR 363-7-4-83-1, an undisturbed clutch of 22 hatched *Troodon* eggs in bottom view from Egg Mountain, Teton County, Montana. Scale bar, 10 cm.

and deposit *en masse*, and incubate within sediments or vegetation mounds^{2,4,6}, whereas birds produce fewer but larger eggs⁷, usually from a single ovary and oviduct³. Further, birds ovulate, shell and lay one egg at a time and incubate eggs directly with body heat³. New discoveries from the Upper Cretaceous of Montana allow re-evaluation of the transition from basal archosaurian to avian reproductive behaviour in the Coelurosauria^{8,9}, the theropod dinosaur clade that includes birds. Egg clutches and nests (Figs 1–3) suggest that the small coelurosaurian *Troodon formosus* (weight, about 50 kg) produced two eggs simultaneously at daily or longer intervals and incubated eggs using a combination of soil and direct body contact. Non-avian coelurosaurians thus possess several primitive features found in crocodilians (two functional ovaries and oviducts, lack of egg rotation and chalazae, partial burial of eggs, precocial young) and several derived features shared with birds (relatively larger and potentially asymmetric eggs, one egg produced per oviduct at a time, loss of egg retention, open nests, brooding) (Fig. 4).

Re-examination of embryonic material shows that eggshell, eggs, egg clutches and nesting horizons^{10–12} previously assigned to the hypsilophodont *Orodromeus*¹³, belong instead to *Troodon*¹⁴. This

study includes these specimens from the Museum of the Rockies (MOR) and two new egg clutches, one associated with a nest structure and the other with an adult *Troodon*. All specimens come from the Campanian Two Medicine Formation of Montana.

Of the existing *Troodon* clutches ($n = 8$), all but two show significant disturbances, such as broken, upended or eroded eggs, or they lack all original matrix owing to preparation procedures. MOR 363-7-4-83-1 contains the bottom two-thirds of 22 hatched eggs. As in other *Troodon* clutches¹¹, the asymmetric eggs stand narrow-end down within the sediments and lean towards the clutch centre. Statistical analysis of egg spacing, trend and plunge within MOR 363 supports a paired arrangement of the eggs (Fig. 1, and see Methods). Similar egg pairing is described in two other coelurosaurians, *Oviraptor*¹⁵ and a possible therizinosaur with large, elongate eggs¹⁶. Neither the mass egg-laying of crocodilians nor the sequential egg-laying of extant birds results in such a pattern. Monoautochronic ovulation, in which each ovary ovulates a single egg simultaneously followed by a period of inactivity, produces pairs of eggs in a variety of lizards¹⁷, and has been observed in an abnormal domestic duck with two functional reproductive tracts¹⁸. No modern examples of pairedness resulting from egg manipulation by adults exists, nor is it suggested by the thermal requirements of eggs. Presence of egg pairs within these theropod clutches implies that there are two functional and avian-like reproductive tracts (in contradiction to ref. 19) with two eggs laid at one- or several-day intervals. Because large clutches require days to weeks to be laid, birds, especially basal lineages with precocial young²⁰, ensure synchronous hatching by delaying incubation until clutch completion³.

Troodon egg shape, size and microstructure also support a more avian than crocodilian reproductive tract. Asymmetric avian egg shape, similar to that of *Troodon*¹² and *Oviraptor*²¹, results from a complicated series of resistances within the oviduct, a process unlikely to occur if more than one egg is present²². Whereas a single 0.5-kg *Troodon* egg is more than ten times larger than eggs of similarly sized extant reptiles, two simultaneously produced *Troodon* eggs are roughly equivalent to the single 1.1-kg egg predicted for a 50-kg bird^{7,12}. Finally, both the angustiprismatic eggshell of *Troodon*^{12,23} and the ornithoid ratite shell of *Oviraptor*²¹ are very avian-like. (Owing to the lack of precise taxonomic information, the phylogenetic distribution of eggshell microstructural features within the Dinosauria remains problematic.) A more

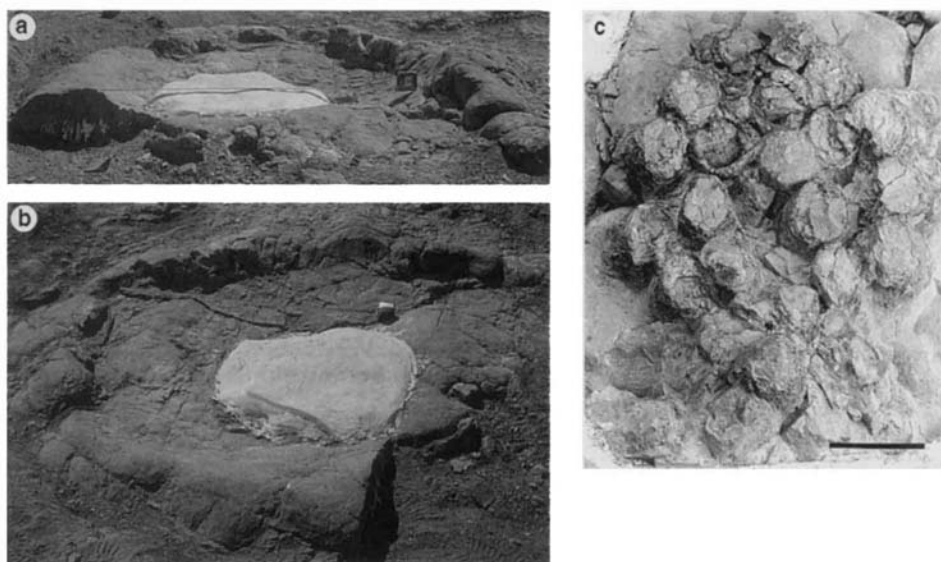


Figure 2 MOR 963, *Troodon* nest with clutch from Egg Mountain. **a, b**, Nest after removal of the soft mudstone fill. Egg clutch sits beneath the white plaster jacket

at the centre. Tape measure corresponds to 1 m. **c**, Clutch of 24 eggs in top view, showing the close proximity of the eggs. Scale bar, 10 cm.

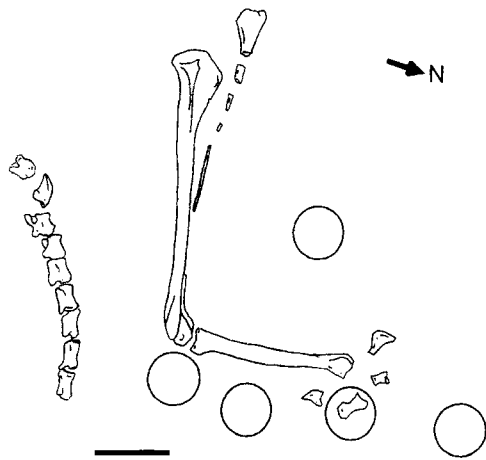


Figure 3 MOR 748, a partial adult *Troodon* preserved in contact with an egg clutch, Glacier County, Montana. An articulated right tibia, fibula, metatarsus, pedal phalanges, a caudal vertebrae series and five eggs are shown. Remaining material (not shown) includes a partial pelvis, right femur, most of the left hindlimb, a series of 8 distal caudals in contact with eggs and at least five eggs. Recent exposure accounts for an indeterminate amount of bone and egg loss. Scale bar, 10 cm.

reptilian oviduct has been inferred for dinosaurs²³ from the presence of a pathological multilayered eggshell. This condition occurs in a number of dinosaur eggshell types²³, but remains unreported in the angustiprismatic and ornithoid eggshell of coelurosaurs.

A new *Troodon* nest (MOR 963) consists of a tight clutch of 24 eggs sitting within a shallow bowl-shaped depression with a distinct rim (Fig. 2). Two lithologies, a muddy micritic limestone and a grey blocky calcareous mudstone comprise the bowl/rim and its fill/cover respectively. X-ray diffraction revealed only slight mineralogical differences between the two lithologies. Nevertheless, the softer mudstone was easily separated using hand tools from the more resistant underlying limestone. Overall, the nest measures 160 × 170 cm externally and ~100 × 100 cm internally. The rim, roughly 10 cm high and 21 cm wide, circles the elliptically shaped clutch ~15–40 cm from the clutch's perimeter. Most egg bottoms sit within the micritic limestone, indicating an unlithified structure at the time of laying. Preparation of this clutch proceeded from top down. The top view shows the tight configuration of eggs resulting from their inward lean, but obscures any egg pairing like that observable in bottom view in the similarly arranged MOR 363 (compare Figs 1 and 2; see Methods).

The *Troodon* nest probably served to house and protect eggs during the lengthy egg-laying period and to incubate the completed clutch. Currently, no evidence for a vegetation cover exists in this or previously described nests. Restriction of eggshell pores to the upper and presumably exposed halves of eggs¹², small clutch area, large nest structure, metabolic indicators²⁴ and preservation of an adult *Troodon* in contact with eggs (Fig. 3), all support the interpretation that *Troodon*, like *Oviraptor*²¹, actively brooded its clutches. The degree of embryonic ossification^{13,25,26} and preservation of intact hatched clutches^{10,11}, indicate that the precocial *Troodon* young left the nest soon after hatching. No evidence suggests the nest served in the post-hatching rearing of young.

The *Troodon* nest differs significantly from the vegetation mounds and hollows of crocodilians² and megapode birds⁵ and more closely approximates the simple open nests common to basal groups of extant birds (Paleognathae, Galliformes and Anseriformes). These birds, like crocodilians and apparently Cretaceous coelurosaurs²⁶, primitively retain self-feeding precocial young which leave the nest shortly after hatching^{6,20,27}. Incubation in

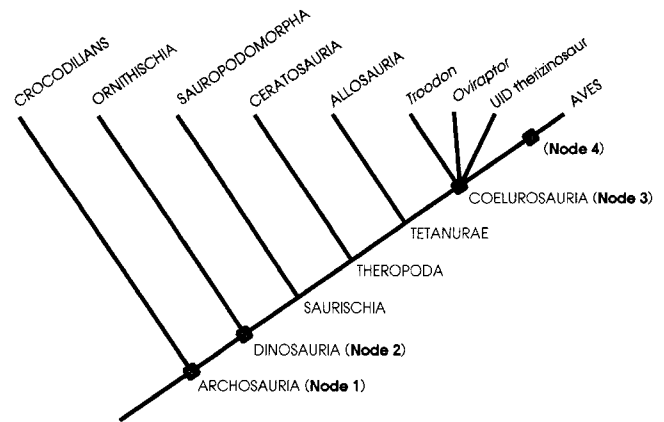


Figure 4 Cladogram of extant archosaurs and selected coelurosaurs based on morphological data²⁸ on which selected reproductive traits are mapped. Node 1: assembly-line oviducts, corpus luteum morphology, hard-shelled eggs, embryonic use of eggshell calcium, clutch attendance, some post-hatching parental care (young attendance). Node 2: organic cores in egg shell units? Node 3: avian-like oviducts, monoautochronic ovulation (two eggs deposited at a time), loss of egg retention, potential for asymmetric eggs, relatively larger egg/adult size ratio, eggs only partially buried, brooding of eggs, delay in incubation from first laying until completion of clutch, open nest. Node 4: loss of right ovary and oviduct function, still larger egg/adult size ratio, eggs free of sediment, egg rotation, chalazae. UID, unidentified.

Troodon and *Oviraptor* appears to be intermediate between crocodilians and birds in using both sediments and direct brooding. *Troodon* eggs stand nearly vertically, half-buried in sediments. Similarly, *Oviraptor* clutches consist of two layers of paired eggs, an arrangement probably requiring at least partial burial^{15,21}. These sediment-bound eggs preclude the possibility of egg rotation as in birds. In crocodilians and other reptiles, rotation of eggs during early development terminates the embryos²⁸. Thus, eggs of non-avian coelurosaurs appear to lack chalazae, the chords of albumen present in birds that allow the embryo to maintain an upright position during egg rotation²⁹. Chalazae and egg rotation probably co-evolved as birds or climbing coelurosaurs sought sediment-free nesting sites such as trees and rocky cliffs. Initially, egg rotation may simply have been an unavoidable consequence of brooding eggs free of sediments. The longer time required by coelurosaurs to generate a clutch with monoautochronic ovulation and brooding may have necessitated a longer pair-bond between mates and greater parental investment in coelurosaurs like *Troodon* in comparison with typical crocodilians. □

Methods

For statistical analysis of clutch MOR 363, the relative locations of eggs were determined by laying a grid (resolution of 0.7 cm) over the clutch photograph and recording the coordinates corresponding to the egg bottoms. Trends and plunges were measured directly using a Brunton compass for all eggs except for one poorly exposed egg. The 22 eggs were paired to minimize the average distance between paired eggs, D , at 6.49 cm. A simulation study then placed 22 eggs with a minimum diameter of 4 cm randomly within a 5-sided polygon representing the clutch area (0.32 m²). Each simulation recognized 11 pairs of eggs that minimized D . 1,000 simulations produced an empirical distribution of $D_{\min}$ values and provided approximate P -values—that is, the probability of observing a $D_{\min}$ at least as extreme as the observed (6.49 cm). For this clutch area, there is strong evidence against a random distribution; P for 6.49 cm equals 0.005. Even for a more conservative $D_{\min}$ value such as 7.5 cm, simulation studies yield $P = 0.051$.

Using the optimal pairings that minimized D , the average difference in trend angle between paired eggs equals 12.1° and the average difference in plunge

angle equals 5.8° . Two permutation-type tests evaluated these figures by generating a large number of random permutations for both the trend and plunge angles and calculating the average difference in trend and plunge angles for each permutation. 10,000 permutations of the trend angles yielded no case with an average difference of less than 12.1° . 10,000 permutations of the plunge angles provided only six cases with an average difference of less than 5.8° . Trend and plunge data also indicate a non-random distribution.

Analysis of MOR 963 followed the same procedures, except that egg tops were used rather than bottoms. The simulation study used a $D_{\min} = 5.3$ cm, a 7-sided polygon with an area of 0.1 m^2 , and yielded $P = 0.342$, neither supporting nor contradicting a paired-egg hypothesis. The top-down preparation of MOR 963 did not permit accurate measurement of egg trends and plunges, nor use of the two permutation-type tests. Egg spacing, trend and plunge data are available from D.J.V.

Received 23 July; accepted 11 November 1996.

1. Packard, G. C., Tracy, C. R. & Roth, J. I. *Biol. Rev.* **52**, 71–105 (1977).
2. Magnusson, W. E., Vliet, A. C., Pooley, A. C. & Magnusson, R. in *Crocodiles and Alligators* (eds Ross, C. A. & Garnett, S.) 118–135 (Facts on Files, New York, 1989).
3. Gill F. B. *Ornithology* (Freeman, New York, 1989).
4. Palmer, B. D. & Guillelme, L. J. *Jr Biol. Reprod.* **46**, 39–47 (1992).
5. Guillelme, L. J. *Jr et al. J. Morph.* **224**, 97–110 (1995).
6. Coombs, W. P. *Jr Spec. Pap. Geol. Soc. Am.* **238**, 21–53 (1989).
7. Blueweiss, L. et al. *Oecologia* **37**, 257–272 (1978).
8. Clark, J. M., Altangerel, P. & Norell, M. A. *Am. Mus. Novit.* **3115**, 1–39 (1994).
9. Holtz, T. R. *Jr J. Paleont.* **70**, 536–538 (1996).
10. Horner, J. R. *Nature* **297**, 675–676 (1982).
11. Horner, J. R. in *Dinosaurs Past and Present II* (eds Czerkas, S. J. & Olson, E. C.) 51–63 (University of Washington Press, Seattle, 1987).
12. Hirsch, K. F. & Quinn, B. *J. Vert. Paleont.* **10**, 491–511 (1990).
13. Horner, J. R. & Weishampel, D. B. *Nature* **332**, 256–257 (1988).
14. Horner, J. R. & Weishampel, D. B. *Nature* **383**, 103 (1996).
15. Dong, Z.-M. & Currie, P. J. *Can. J. Earth Sci.* **33**, 631–636 (1996).
16. Currie, P. J. *Natl Geogr. Mag.* **189**, 96–111 (1996).
17. Smith, H. M., Sinelnik, G., Fawcett, J. D. & Jones, R. E. *Trans. Kans. Acad. Sci.* **75**, 107–120 (1973).
18. Kinsky, F. C. *J. Ornithol.* **112**, 334–357 (1971).
19. Jones, R. E., Smith, H. M. & Bock, C. E. *J. Theor. Biol.* **161**, 527–536 (1993).
20. Sillen-Tullberg, B. & Temrin, H. in *Phylogenetics and Ecology* (eds Eggleton, P. & Vane-Wright, R. I.) 311–322 (Proc. R. Soc., London, 1994).
21. Norell, M. A., Clark, J. M., Chiappe, L. M. & Dashzeveg, D. *Nature* **378**, 774–776 (1995).
22. Iverson, J. B. & Ewert, M. A. in *Egg Incubation* (eds Deeming, D. C. & Ferguson, M. W. J.) 87–100 (Cambridge University Press, Cambridge, 1991).
23. Hirsch, K. F. in *The Paleobiology of Trace Fossils* (ed. Donovan, S. K.) 269–294 (Johns Hopkins, Baltimore, 1994).
24. Varricchio, D. J. *J. Vert. Paleont.* **13**, 99–104 (1993).
25. Stark, J. M. *J. Morphol.* **222**, 113–131 (1994).
26. Geist, N. R. & Jones, T. D. *Science* **272**, 712–714 (1996).
27. Sibley, C. G. & Ahlquist, J. E. *Phylogeny and Classification of Birds* (Yale University Press, New Haven, 1990).
28. Deeming, D. C. in *Egg Incubation* (eds Deeming, D. C. & Ferguson, M. W. J.) 307–323 (Cambridge University Press, Cambridge, 1991).
29. Terres, J. K. *The Audubon Society Encyclopedia of North American Birds* (Wings Books, New York, 1995).

Acknowledgements: We thank the Museum of the Rockies, Montana State University, the Old Trail Museum, M. Lange and the rest of the MOR's 1994 Field School participants, P. Leiggi, Y. Katsura, T. Weaver, C. Trueman, the families of V. and L. Carrol, E. Bray, R. Jackson, K. Carpenter, K. Alf, G. Erickson and G. Dyke; we thank P. Currie, K. Padian, R. Rogers and H. Temrin for critically reading early drafts of the manuscript. B. Selyem (MOR) provided egg photos.

Correspondence and requests for materials should be addressed to D.J.V. (e-mail: otmdjv@3rivers.net).

Body-size evolution in Cretaceous molluscs and the status of Cope's rule

David Jablonski

Department of Geophysical Sciences, University of Chicago,
5734 South Ellis Avenue, Chicago, Illinois 60637, USA

Cope's rule, the tendency for lineages to evolve to larger body size, is widely seen as a pervasive evolutionary pattern^{1–4}. However, only a few studies^{5–8} have gone beyond enumerating isolated examples to assess its overall frequency relative to body-size decrease or stasis. Thus, although size is clearly an important parameter for microevolution and ecology^{9–14}, including conservation biology¹⁰, its impact on large-scale patterns remains poorly

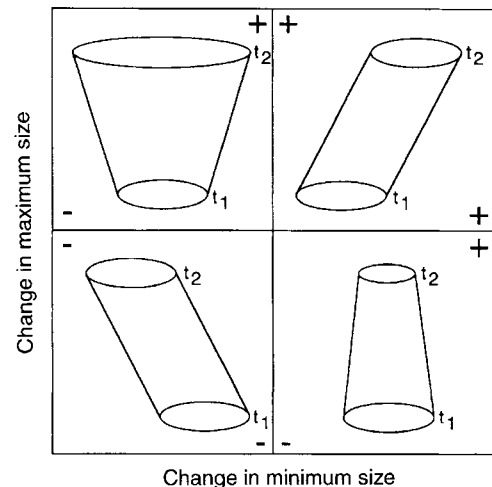


Figure 1 A graphical approach to the analysis of evolutionary changes in body size, with each quadrant representing a different temporal pattern, as shown schematically. (t_1 = time 1; t_2 = a subsequent time 2.) The vertical axis is the change in the upper bound of the size distribution of the species in a lineage, and the horizontal axis is the change in the lower bound of the size distribution. Each clade can thus be plotted as a point determined by the behaviour of its upper and lower size bounds (in this study, the maximum recorded size of the largest and smallest species, respectively). The top-right quadrant represents Cope's rule, the bottom-left a directional evolutionary size decrease, the top-left an increase in the range of body sizes or 'increase in variance', and the bottom-right a decrease in size range.

understood. The prevalence of Cope's rule is even more uncertain, as some reported cases of evolutionary size increase may actually represent an expansion of a clade's size range (a pattern generally termed an 'increase in variance', although not necessarily in the formal statistical sense) rather than a phyletic, directional trend^{15–18}. I have performed a comprehensive census of body-size changes in a large fauna of Cretaceous bivalve and gastropod genera. A directional net increase in body size (including the loss of small-sized species and thus representing Cope's rule in the strict sense) is no more frequent than an increase in size range among species or a net evolutionary size decrease. Thus the undisputed ecological importance of body size does not translate into a preferred macroevolutionary pattern.

The Late Cretaceous (Late Santonian–Maastrichtian) deposits of the Gulf and Atlantic Coastal Plain of North America contain a rich and well-preserved marine molluscan fauna¹⁹. This period of 16 Myr, which can be divided into increments of 2 Myr, contains 191 bivalve and gastropod lineages that are generally afforded genus or subgenus rank²⁰. (The genus is a standard level of analysis for body-size trends, as discussions of Cope's rule have almost always involved claims about large-scale patterns^{8,16,21}; genera and subgenera are here treated equivalently as operational units because the distinction between the two ranks is often arbitrary.) Several apparently paraphyletic taxa were amalgamated for this analysis, but most of the genera and subgenera originated before the study interval, thereby reducing the number of potential paraphyly-derived artefacts.

The data consist of adult sizes for each genus at the beginning and the end of the study interval, as determined by the geometric mean of height and length²² of the largest known specimen of each constituent species. Sizes were obtained for all described species and a number of undescribed ones (total $n = 1,086$), but the rudist bivalves were excluded as they were considered too rare and fragmentary in this region for meaningful measurement. Maximum sizes were used to exclude immature specimens (most of these species exhibit indeterminate growth²³), but results are not qualitatively different when median sizes are taken for groups of probably

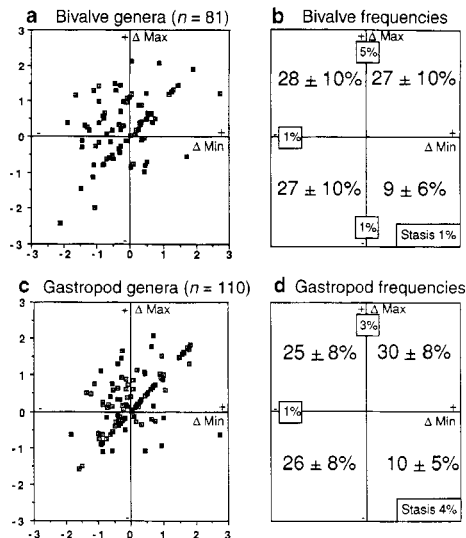


Figure 2 Evolutionary patterns of size change in genera and subgenera of Late Cretaceous bivalves (**a**, **b**) and gastropods (**c**, **d**) over 16 Myr. Sizes were $\log_2$ -transformed, so that, for example, an increase of unit 1 represents a doubling, and a corresponding decrease a halving, of body size; 95% confidence limits on frequencies calculated following ref. 27. ΔMax = change in upper bound of adult sizes; ΔMin = change in lower bound of adult sizes.

adults in the 50 best-sampled genera. Because shapes are relatively constant within genera, and because many causal explanations involve volume-correlated parameters (such as gonad output, physiological inertia, size refuges from predation), more elaborate measures such as the first principal component of a multivariate morphometric analysis were not used. Analysis at this taxonomic level may lose details of individual species transitions (but see below for data on phyletic trends), but it allows the partitioning of large-scale trends according to net directional or diffusional dynamics in a large number of discrete phylogenetic units.

The evolution of body size can be analysed using the scheme shown in Fig. 1, with each taxon plotted as a point determined by the change in both extremes of its size range, as defined by the largest and smallest species. Taking this approach, the bivalve and gastropod genera show similar patterns of body-size evolution, with statistically equivalent frequencies of net directional size increase (27–30%), net size decrease (26–27%), and increased variance (25–28%) (Fig. 2). The paucity of taxa exhibiting either a decrease in variance or stasis conforms to the expectation if body sizes commonly evolve within genera as a diffusion process. Only 3–5% of the taxa exhibit stasis at the lower bound while expanding towards greater sizes; this suggests that few of these molluscs are near a clade-specific or general size limit, which can impose apparent directionality on diffusive or passive trends^{15,18}. (By tallying taxa according to the loss or accretion of species at their lower as well as upper size bounds, these analyses essentially correspond to McShea's¹⁸ "test based on the behavior of the minimum" for distinguishing passive from active, or directional, trends.)

The alignment of taxa on the diagonal in the upper right and lower left quadrants is in a sense an artefact of the analytical technique. Those points represent lineages that were monospecific in both their first and last time increment; the lone species is by default both smallest and largest in its time interval, so that any net size increase or decrease inevitably plots as a symmetrical change with respect to both minimum and maximum size. This subset of the data, which provides one estimate of the relative frequencies of phyletic size change, also shows no significant tendency towards Cope's rule ($16 \pm 8\%$ of all bivalve genera show this phyletic size increase, $14 \pm 8\%$ show size decrease; $18 \pm 7\%$ of all gastropod genera show phyletic size increase, $15 \pm 7\%$ show size decrease).

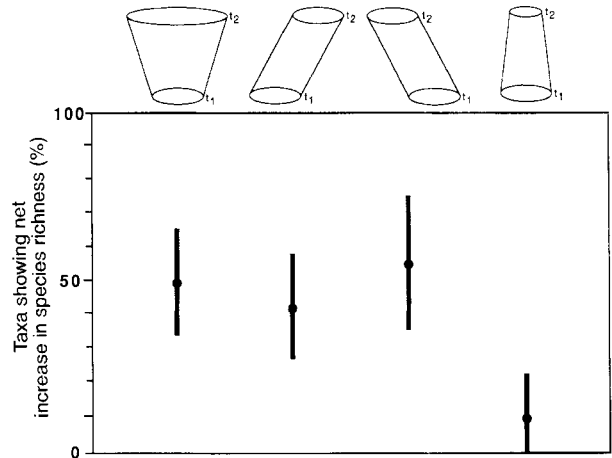


Figure 3 Percentage of taxa exhibiting a net increase in species richness. The increased-variance category is not significantly dominated by taxa showing a net increase in species numbers, relative to taxa showing directional size trends (G-tests on raw data, with Williams' correction). This suggests that the increased-variance pattern is not simply an artefact of sampling more species late in the study interval. The decreased-variance pattern is significantly concentrated in clades with declining species richness, as expected if body sizes evolve as a random walk. Taxa constrained by monospecific start and finish are omitted, as are those in which the upper and/or lower bound is static; 95% confidence limits on frequencies calculated following ref. 27.

If taxa that fall on the diagonal are omitted, then the 'increased variance' quadrant is most heavily occupied (39% and 40% of the remaining genera of bivalves ($n = 57$) and gastropods ($n = 70$), respectively), and the quadrants presenting Cope's rule and net size decrease still have equal frequencies (17–18% and 18–20%, respectively). As before, an evolutionary decrease in variance is less common, although no longer significantly so (11–16%), and stasis is seen least of all. Again, this is the pattern expected for a large population of clades branching stochastically to larger and smaller sizes. The slightly greater proportion of points in the upper half of Fig. 2a ($60 \pm 10\%$) and Fig. 2c ($58 \pm 9\%$) derives from the rarity of 'decreased variance' clades relative to those that increase their size range or show size trends upwards or downwards, rather than from any overall tendency towards Cope's rule.

These results are unlikely to be artefacts of taxonomic practice or the geographic limits of the study. Incorporation of extraprovincial species, or correction of any remaining paraphyletic taxa by merging them into monophyletic units, would be as likely to create an increase or decrease in size range as to generate a directional pattern. A pattern typical of Cope's rule, for example, would be produced only by the addition of a smaller species at the start of a formerly increased-variance clade, or by the addition of a larger species at the end of a formerly decreased-variance clade. Conversely, Cope's rule would be converted to decreased variance by the addition of a larger species at the start of the study interval, or to increased variance by the addition of a smaller species at the end of the study interval.

The youngest time increment, the *Haustator bilira* biostratigraphic zone, has more localities than the oldest time increment, and this difference might artificially increase size ranges within a genus through time. However, when the three oldest time increments of the study interval are pooled to raise the number of localities to about 150, as in the youngest time increment, the resulting changes in frequency are statistically negligible—frequencies in the upper right quadrant (i.e. Cope's rule), for example, are $32 \pm 10\%$ and $30 \pm 8\%$, for bivalves and gastropods respectively—suggesting that sampling does not create the overall pattern. Indeed, by summing the evolutionary spread of body sizes in the initial intervals, this procedure must artificially reduce the number of taxa

falling into the increased-variance quadrant. Further, the proportion of clades exhibiting a net increase in species richness is no higher for the increased-variance group than for the directional-change groups (Fig. 3), so the observed patterns are again unlikely to be sampling artefacts. By indicating that patterns are not driven by changes in sample size, these tests suggest that the observed expansions or contractions in size range represent true changes in variance.

The analyses presented here do not support Cope's rule as an evolutionary generalization, and provide the most extensive empirical evidence yet for the 'increased variance' pattern as an equally important pathway in body-size evolution. The selective forces often held to drive Cope's rule (for example, advantages of large size in defence, mating success, predatory ability, and resistance to environmental extremes^{1-4,21,26}) generally imply the evolutionary loss of small-bodied forms. However, the fossil record shows that increases in the maximum body size within clades do not require such mechanisms, but can simply represent one limb of an expanding size range, and that directional decreases in sizes are no less frequent than directional increases. This apparently random evolution of body size in Cretaceous molluscs need not imply the absence of driving mechanisms, but could result from the interaction of many factors which have effects so context specific and scale dependent that unitary patterns fail to emerge¹⁸. Large size is not universally advantageous, and multiple pressures operate on body size and taxon-specific correlates that range from age at first reproduction to allometric morphologies^{13,16}. Thus lineages collectively fail to follow a single, predictable size trajectory as their species diffuse or shift through size changes. Extinction probability also need not be strongly related to size^{24,25}. Despite its importance in microevolution and ecology⁹⁻¹⁴, size plays a surprisingly weak or unpredictable role at larger scales, reinforcing the view that macroevolutionary patterns need not be simple extensions of those seen at the level of individual organisms over microevolutionary time. □

Received 13 May; accepted 23 October 1996.

1. Bonner, J. T. *The Evolution of Complexity* (Princeton Univ. Press, 1988).
2. Benton, M. J. in *Paleobiology: A Synthesis* (eds Briggs, D. E. G. & Crowther, P. R.) 147–152 (Blackwell, Oxford, 1989).
3. Hallam, A. in *Causes of Evolution* (eds Ross, R. M. & Allmon, W. D.) 249–269 (Univ. Chicago Press, 1990).
4. Bush, G. L. in *Evolutionary Patterns and Processes* (eds Lees, D. R. & Edwards, D.) 230–249 (Academic, London, 1993).
5. MacFadden, B. J. *Fossil Horses* (Cambridge Univ. Press, 1992).
6. Norris, R. D. *Paleobiology* 17, 388–399 (1991).
7. Arnold, A. J., Kelly, D. C. & Parker, W. C. *J. Paleontol.* 69, 203–210 (1995).
8. Gingerich, P. D. *Annu. Rev. Earth Planet. Sci.* 8, 407–424 (1980).
9. McLain, D. K. *Oikos* 68, 490–500 (1993).
10. Pimm, S. L. *The Balance of Nature?* (Univ. Chicago Press, 1991).
11. Maurer, B. A., Brown, J. H. & Rusler, R. *Evolution* 46, 939–953 (1992).
12. Brown, J. H. *Macroecology* (Univ. Chicago Press, 1995).
13. LaBarbera, M. in *Patterns and Processes in the History of Life* (eds Raup, D. M. & Jablonski, D.) 69–98 (Springer, Berlin, 1986).
14. Peters, R. H. *The Ecological Implications of Body Size* (Cambridge Univ. Press, 1983).
15. Stanley, S. M. *Evolution* 27, 1–26 (1973).
16. McKinney, M. L. in *Evolutionary Trends* (ed. McNamara, K. J.) 75–118 (Univ. Arizona Press, Tucson, 1990).
17. Gould, S. J. *J. Paleontol.* 62, 319–329 (1988).
18. McShea, D. W. *Evolution* 48, 1747–1763 (1994).
19. Jablonski, D. & Raup, D. M. *Science* 268, 389–391 (1995).
20. Jablonski, D. *Bull. Mar. Sci.* 39, 565–587 (1986).
21. Newell, N. D. *Evolution* 3, 103–124 (1949).
22. Stanley, S. M. *Paleobiology* 12, 89–110 (1986).
23. Vermeij, G. J. & Signor, P. W. *Biol. J. Linn. Soc.* 47, 233–247 (1992).
24. Jablonski, D. in *Evolutionary Paleobiology* (eds Jablonski, D., Erwin, D. H. & Lipps, J. H.) (Univ. Chicago Press, in the press).
25. Gaston, K. J. & Blackburn, T. M. *Phil. Trans. R. Soc. Lond. B* 347, 205–212 (1995).
26. Brown, J. H. & Maurer, B. A. *Nature* 324, 248–250 (1986).
27. Raup, D. M. in *Analytical Paleontology* (eds Gilinsky, N. L. & Signor, P. W.) 207–216 (Paleontological Society, Knoxville, TN, 1991).

Acknowledgements: I thank S. Calzada, R. J. Cleavelly, A. V. Dhondt, D. L. Dockery, C. Heinberg, C. C. Johnson, E. G. Kauffman, G. R. Kennedy, M. Machalski, N. J. Morris, L. R. Saul, N. F. Sohl and G. L. Wingard for access to collections and unpublished information; J. Walsh for assistance with data entry; and W. D. Allmon, M. J. Benton, D. H. Erwin, M. Foote, S. J. Gould, S. M. Kidwell, B. A. Maurer, F. K. McKinney, A. I. Miller, P. Pearson, K. Roy, S. M. Stanley, P. J. Wagner and P. D. Ward for comments. This work was supported by the US National Science Foundation.

Correspondence and requests for materials should be addressed to the author (e-mail: djablons@midway.uchicago.edu).

Habitat heterogeneity as a determinant of mammal species richness in high-energy regions

Jeremy T. Kerr & Laurence Packer

Department of Biology, York University, 4700 Keele Street, North York, Ontario M3J 1P3, Canada

A fundamental problem in ecological research is to explain large-scale gradients in species richness^{1,2}. Although many causative agents for this phenomenon have been suggested, the species richness–energy hypothesis has received the strongest empirical support^{3–6}: this hypothesis states that higher energy availability provides a broader resource base, permitting more species to coexist. Here we show that the species richness–energy hypothesis applies to North American mammals only over a limited geographical area in which climatic energy levels are low (Alaska and most of Canada), rather than on a continental scale as had previously been accepted⁶. In relatively high-energy regions of North America, corresponding to most of the continental United States and southern Canada, we find that mammal species richness is best predicted by topographic heterogeneity and local variation in energy availability. Our results contradict previous studies of large-scale richness patterns that dismissed the importance of habitat heterogeneity^{7–9}, and have implications for climate change research.

Climatic factors, environmental stability, land area, habitat heterogeneity, historical influences (such as Pleistocene glaciations) and energy availability are the factors most often discussed as determinants of regional variability in species richness^{8–13}. Energy

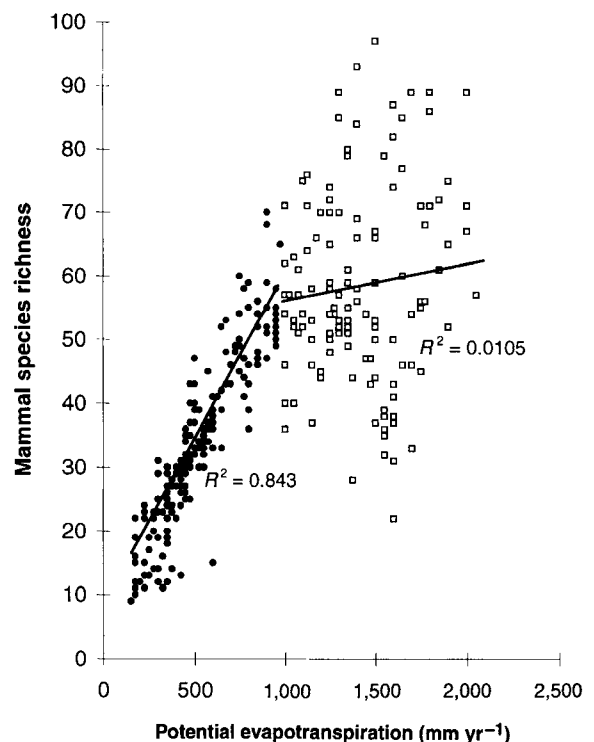


Figure 1 The relationship between PET and mammal species richness in North America. In areas where PET < 1,000 mm yr⁻¹, PET explains 84% of the variance in mammal richness ($F = 1.096$, $P \ll 0.0001$). South of this zone, however, PET is unrelated to mammal richness ($F = 1.35$, $P = 0.248$).

Table 1 Correlations between environmental factors and mammal species richness in regions in North America

Environmental variable per quadrat ⁶	Pearson correlation with MSR† in regions where PET $\geq 1,000 \text{ mm yr}^{-1}$	Pearson correlation with MSR‡ in regions where PET $< 1,000 \text{ mm yr}^{-1}$
Mean annual temperature (1)	-0.392***	0.837***
Mean PET (1)	NS	0.920***
Mean actual evapotranspiration (1)	-0.784***	0.709***
Mean solar radiation (1)	0.353***	0.789***
Mean precipitation (2)	-0.522***	0.318***
Elevation variability (3)	0.808***	0.300***
Precipitation variability (3)	0.344***	0.177*
PET variability (3)	0.615***	0.542***
Annual temperature variability (4)	NS	-0.373***
Glaciation (5)	NS	NS
Longitude	0.698***	-0.150*
Latitude	0.335**	-0.806***
Quadrat area	0.194*	0.238*
Coastal location	-0.176*	-0.571***
Peninsular location	-0.215*	-0.204**

The number after the environmental variable refers to the hypothesis that the variable tests (see Methods). Coastal and peninsular location and quadrat area are control variables; latitude and longitude provide spatial reference. MSR, mammal species richness; PET, potential evapotranspiration. * $P < 0.05$; ** $P < 0.005$; *** $P < 0.0001$; NS, not significant.

† $n = 130$.

‡ $n = 206$.

explains most of the observed variance in regional species richness patterns (median R^2 value of 70%, based on 41 studies¹⁴). Potential evapotranspiration (PET, the amount of water that would evaporate from a saturated surface), an aspect of climatic energy availability, is the best predictor of richness patterns among North American vertebrates, explaining 80% of the geographical variation in mammal species richness⁶. However, inspection of Fig. 1, combined with split-line regression techniques, shows that potential evapotranspiration is closely related to mammal richness when PET $< 1,000 \text{ mm yr}^{-1}$, but not south of this region (Figs 1 and 2). We therefore investigated other predictors of mammal species richness for higher energy areas south of the PET isocline of $1,000 \text{ mm yr}^{-1}$, an area where mammal richness patterns and latitude do not co-vary¹⁵ (Table 1).

Two aspects of habitat heterogeneity, topographical variation and spatial variability in potential evapotranspiration, emerge from stepwise regression analysis as primary predictors of mammalian species richness (Fig. 3). The high rate at which habitats change along an elevational gradient produces high between-habitat diversity in regions with greater topographic variability^{16,17}, leading to

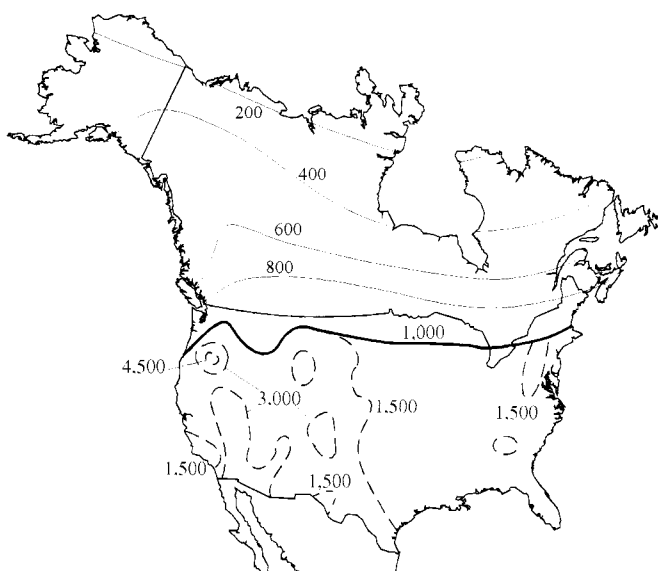


Figure 2 A map of North America showing PET patterns (solid curves) in the north (mm yr^{-1}) and topographical heterogeneity (dashed curves) in the south (metres). The PET contour at $1,000 \text{ mm yr}^{-1}$ is in bold. North of this contour, PET is the best predictor of mammal richness, whereas heterogeneity predicts richness to the south.

increased regional species richness. Variability in potential evapotranspiration within quadrats may influence mammal species richness through the large-scale co-occurrence of species adapted to different levels of energy availability.

Our results indicate that although there is no single determinant of large-scale variation in mammal species richness, there may be a hierarchical sequence of limiting factors. Our data are consistent with the hypothesis that energy availability limits richness at high latitudes, but that habitat heterogeneity is important when

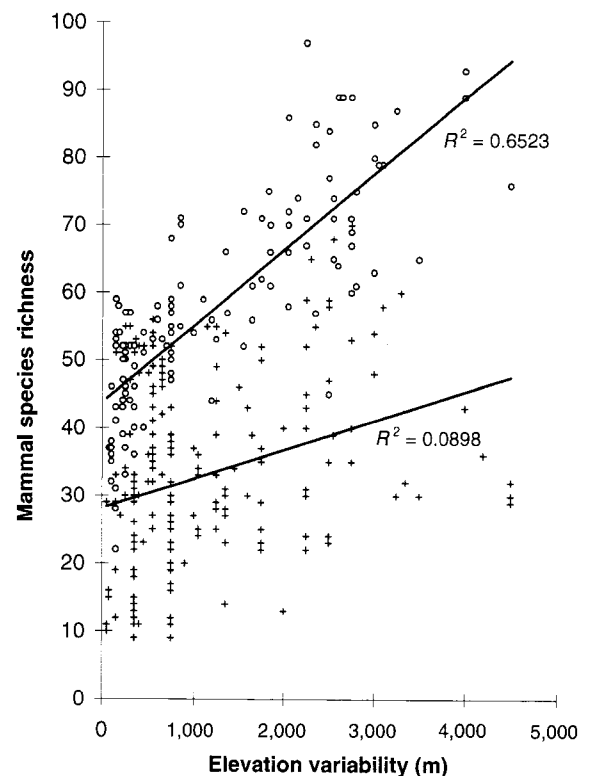


Figure 3 The contrasting relationships between mammal species richness and elevation variability, depending on PET levels (crosses when PET $< 1,000 \text{ mm yr}^{-1}$, and open circles when PET $\geq 1,000 \text{ mm yr}^{-1}$). In combination with PET variability and coastal location, these factors explain 76.7% of the variability in mammal species richness patterns in high-energy regions of North America ($F = 138.4$, $P < 0.0001$; mammal species richness = $43.3 + 0.00852 \times$ topographical heterogeneity + 0.0354 PET variability - 9.60 coastal location). In regions where PET $< 1,000 \text{ mm yr}^{-1}$, elevation variability is a poor predictor of mammal richness ($F = 20.12$, $R^2 = 0.0898$, $P < 0.0001$).

PET $\geq 1,000 \text{ mm yr}^{-1}$. Other factors may predict mammal species richness in tropical regions where mammal distributions are relatively poorly known. As potential evapotranspiration in the American southwest is comparable to that of the Amazonian basin¹⁸, we have sampled almost the entire range of global variation and found that energy is important only in comparatively cold regions.

The nonlinear relationships between diversity and energy availability that we have described may have significant implications for the impacts of climate change on mammal communities. As energy availability increases because of global warming, habitat heterogeneity may become the prime determinant of species richness patterns over an increasing proportion of the North American continent. □

Methods

Description of data. We investigated predictors of mammal species richness in North America using data from ref. 6. We determined the approximate PET contour beyond which PET is unrelated to mammal richness by visual inspection of the PET-richness plot (Fig. 1) and general agreement between the Quasi-Newton, Simplex and Hooke–Jeeves breakpoint estimation routines¹⁹. The high energy region consists of $130 \text{ } 2.5^\circ \times 2.5^\circ$ quadrats. Independent variables describing mean conditions were determined by averaging maximum and minimum values for the different environmental variables in each quadrat. Those measuring spatial variability were determined by taking the difference between the maximum and minimum values per quadrat for the respective environmental descriptors. Annual temperature variability is the difference between the mean January and July temperatures, respectively, of each quadrat. Glaciation effects were measured by creating a dummy variable describing whether quadrats were clear, inundated or glaciated during the Wisconsinian. Coastal and peninsular location are also dummy variables. Coastal location, in particular, accounts for low richness outliers in our first and third figures.

Statistical analysis. We analysed bivariate plots of species richness and the various predictor variables (Table 1) in this region and tested the observed relations using both forward and backward stepwise regression analysis^{19,20}. Quadrat area does not enter our final model. Using correspondingly numbered variables in Table 1, we investigated the following hypotheses of species diversity: (1) species richness–energy^{21–23}; (2) climatic favourability³; (3) habitat heterogeneity¹⁵; (4) climatic stability²⁴; and (5) glacial history¹¹. Variation in both elevation and PET are consistently the most important predictors of mammal richness, regardless of the regression approach used. Additional variables may be added to the regression equation, but contribute little to the predictive power of the model.

Received 12 August; accepted 27 November 1996.

1. Andrewartha, H. G. *Introduction to the Study of Animal Populations* (University of Chicago Press, Chicago, 1961).
2. Rohde, K. *Oikos* **65**, 514–527 (1992).
3. Currie, D. J. & Paquin, V. *Nature* **329**, 326–327 (1987).
4. Turner, J. R. G., Gatehouse, C. M. & Corey, C. A. *Oikos* **48**, 195–205 (1989).
5. Turner, J. R. G., Lennon, J. J. & Lawrenson, J. A. *Nature* **335**, 539–541 (1989).
6. Currie, D. J. *Am. Nat.* **137**, 27–49 (1991).
7. Stevens, G. C. *Am. Nat.* **133**, 240–256 (1989).
8. Schall, J. J. & Pianka, E. R. *Science* **201**, 679–686 (1978).
9. Adams, J. M. & Woodward, F. I. *Nature* **339**, 699–701 (1989).
10. Wallace, A. R. *Tropical Nature and Other Essays* (Macmillan, New York, 1878).
11. Pianka, E. R. *Am. Nat.* **100**, 33–46 (1966).
12. Huston, M. *Am. Nat.* **113**, 81–101 (1979).
13. Rosenzweig, M. L. *Species Diversity in Space and Time* (Cambridge University Press, New York, 1995).
14. Wright, D. H., Currie, D. J. & Maurer, B. A. *Species Diversity in Ecological Communities* (eds Ricklefs, R. E. & Schluter, D.) 66–76 (University of Chicago Press, Chicago, 1993).
15. Simpson, G. G. *Syst. Zool.* **13**, 57–73 (1964).
16. World Conservation Monitoring Centre *Global Biodiversity: The Status of Earth's Living Resources* (Chapman & Hall, New York, 1992).
17. Rahbek, C. *Ecography* **18**, 200–205 (1995).
18. USSR Committee for the International Hydrological Decade *Atlas of World Water Balance* (UNESCO, Paris, 1977).
19. StatSoft, Inc. *STATISTICA for Windows* (Statsoft Inc., Tulsa, 1995).
20. Zar, J. H. *Biostatistical Analysis* (Prentice Hall, Toronto, 1984).
21. Hutchinson, G. E. *Am. Nat.* **93**, 145–159 (1959).
22. Connell, J. H. & Orias, E. *Am. Nat.* **98**, 399–414 (1964).
23. Wright, D. H. *Oikos* **41**, 496–506 (1983).
24. Klopfer, P. H. *Am. Nat.* **93**, 337–342 (1959).

Acknowledgements: We are grateful to the Natural Sciences and Engineering Research Council of Canada for scholarship support for J.T.K. and research funding for L.P. We thank D. J. Currie for discussion and for permission to use his climatic and mammal-richness data, M. De Giusti for technical support, and L. Atkinson, B. Fenton and J. Shore for helpful suggestions.

Correspondence and requests for materials should be addressed to J.T.K. (e-mail: jtkerr@yorku.ca).

Impaired auditory recognition of fear and anger following bilateral amygdala lesions

Sophie K. Scott*, Andrew W. Young*, Andrew J. Calder*, Deborah J. Hellawell†, John P. Aggleton‡ & Michael Johnson§

* MRC Applied Psychology Unit, 15 Chaucer Road, Cambridge CB2 2EF, UK

† Rehabilitation Studies Unit, Astley Ainslie Hospital, Grange Loan, Edinburgh EH9 2HL, UK

‡ School of Psychology, University of Wales at Cardiff, PO Box 901, Cardiff CF1 3YG, UK

§ Department of Neurology, St James's University Hospital, Beckett Street, Leeds LS9 7TF, UK

The amygdalar complex is a medial temporal lobe structure in the brain which is widely considered to be involved in the neural substrates of emotion. Selective bilateral damage to the human amygdala is rare, offering a unique insight into its functions. There is impairment of social perception after amygdala damage, with defective recognition of facial expressions of emotion^{1–4}. Among the basic emotions, the processing of fear and anger has been shown to be disrupted by amygdala damage^{1,2,5}. Although it remains puzzling why this not found in all cases⁶, the importance of the amygdala in negative emotion, and especially fear, has been confirmed by conditioning⁷, memory⁸ and positron emission tomography (PET) experiments^{9,10}. Central to our understanding of these findings is the question of whether the amygdala is involved specifically in the perception of visual signals of emotion emanating from the face, or more widely in the perception of emotion in all sensory modalities¹¹. We report here a further investigation of one of these rare cases, a woman (D.R.) who has impaired perception of the intonation patterns that are essential to the perception of vocal affect, despite normal hearing. As is the case for recognition of facial expressions, it is recognition of fear and anger that is most severely affected in the auditory domain. This shows that the amygdala's role in the recognition of certain emotions is not confined to vision, which is consistent with its being involved in the appraisal of danger and the emotion of fear^{12,13}.

D.R., a woman in her early fifties, first suffered from epilepsy at the age of 28. After anticonvulsant drugs failed to control this, she underwent a series of stereotaxic operations targeted at the left and right amygdala. Tracings of the lesions in the region of the amygdala from magnetic resonance imaging (MRI) scans are shown in Fig. 1.

Following surgery, D.R. could recognize the faces of people who were familiar to her before the operation, and she performed well on unfamiliar face-matching tasks in which the faces had neutral expressions. Hence, there was no general impairment of face perception. In contrast, D.R. showed poor processing of social signals from the face; her judgment of direction of gaze and her interpretation of facial expressions of emotion were impaired^{3,4}. For facial expressions, recognition of fear was differentially severely affected, but there was also evidence of impaired recognition of anger and (to a lesser extent) disgust⁵.

For the present study, we assessed D.R.'s perception of sounds and voices. Audiometric testing was unremarkable, with hearing within normal limits for her age for all frequencies tested (250–8,000 Hz). Tests from the PALPA battery¹⁴ were used to examine D.R.'s auditory processing of language: the results are shown in Table 1. D.R. had no significant problems in discriminating minimally different pairs of

Table 1 Performance of D.R. on auditory and voice perception tasks

	D.R.	Controls	
		Mean	s.d.
PALPA auditory processing tests			
Discrimination of non-word minimal pairs			
Same	36/36	35.70	0.56
Different	32/36	35.09	2.34
Discrimination of word minimal pairs			
Same	35/36	35.54	0.78
Different	35/36	34.83	2.58
Phonological segmentation of initial sounds			
Words	29/30	29.36	0.86
Non-words	13/15	14.24	1.01
Recognition of environmental sounds			
	15/20	16.85	1.42
Vocal expressions			
Sentence content control task			
Tone of voice	21/24	23.15	1.35
	17/24**	21.50	1.70
Sentence prosody	9/24***	19.90	3.40
Unfamiliar voice matching			
	27/40***	36.75	2.61
Recognition of familiar voices			
Famous voices			
Recognized as familiar	6/30**	20.30	5.65
Correctly identified	2/30**	15.70	5.88
Unfamiliar voices			
Correct rejections	8/10	9.20	0.77

Control data for the PALPA subtests are taken from the test's norms¹⁴; from the other tests, control data come from ref. 15. Asterisk scores are significantly below the control mean: * $z > 1.65$, $P < 0.05$; ** $z > 2.33$, $P < 0.01$; *** $z > 3.10$, $P < 0.001$.

spoken nonwords (such as 'zog-zog' vs. 'zog-zeg') or words (such as 'house-house' vs. 'house-mouse'), and she was able to identify the initial phonemes from monosyllabic words and non-words.

We also tested D.R.'s ability to identify environmental sounds, to interpret vocal expressions of emotion, to match unfamiliar voices, and to recognize familiar people from their voices. Full details of these tasks are given elsewhere¹⁵; D.R.'s performance is summarized in Table 1.

To test recognition of environmental sounds, we used 20 clips of everyday sounds (such as rain or the telephone) lasting 8 s each. D.R. was asked to identify each sound. Her performance fell in the low-normal range (15/20 correct; $z = 1.30$, $P > 0.05$).

For our present purposes, the tasks summarized so far show that any problems that D.R. had in recognizing auditory emotion could not be due to generalized hearing defect. In addition, we established that she understood the meaning of 'emotion' words: D.R. could readily describe circumstances in which other people would experience emotions of happiness, sadness, anger, and so on. To test this formally, a control task was given, in which D.R. was asked whether the content of 24 sentences was happy, angry or sad. She again performed in the low-normal range (21/24 correct; $z = 1.59$, $P > 0.05$), showing adequate comprehension of the meanings of emotional terms.

To assess D.R.'s ability to understand vocal expressions of emotion, she was asked to listen to a tape of 24 sentences with neutral content. These sentences were read with an emotional tone of voice, and D.R. was asked whether each sentence was read in a happy, angry or sad manner. Her performance at identifying emotion from the tone of voice was well below that of the controls (17/24 correct; $z = 2.65$; $P < 0.01$).

Our unfamiliar voice-matching task involved listening to a taped set of 40 trials, on each of which the same sentence was read out twice. For half of these trials, the sentence was read by different people, and for the other half it was read by the same person. D.R.'s performance was very poor (27/40 correct overall; $z = 3.74$, $P < 0.001$) in deciding whether each sentence had been read by the same or different people. A further sentence prosody task involved 24 sentences with neutral content, read with intonation

patterns indicating a question, a statement or an exclamation: D.R. was severely impaired at deciding what each sentence was (9/24 correct; $z = 3.21$; $P < 0.001$), performing at chance level (chance = 8/24 correct).

Recognition of familiar voices was tested using 30 famous people's voices and 10 unfamiliar voices, recorded in clips of about 5 s long. D.R. was asked whether each voice was familiar, and if so to identify the person by occupation or name. She performed poorly both at recognizing voices as familiar (6/30 correct; $z = 2.53$, $P < 0.01$) and identifying them (2/30 correct; $z = 2.33$, $P < 0.01$).

Across these tasks, then, D.R. had impaired ability to make use of intonation patterns with communicative significance, including those used in signalling emotion¹⁶. It is likely that D.R.'s problems in matching unfamiliar voices and recognizing familiar voices also derived from the same source, because intonation patterns form an important determinant of what voices sound like^{17,18}.

To explore further D.R.'s problems in auditory recognition of emotion, we developed two tasks with enough trials to allow assessment of different basic emotions; data are summarized in Table 2.

The first test involved single words with neutral meanings (for example, carpet), spoken with intonations intended to convey emotions of happiness, sadness, anger, fear or disgust. These emotions were chosen because they are each known to be associated with distinct sets of prosodic features¹⁹. D.R. showed significantly impaired recognition of anger (3/10 correct; $z = 4.56$; $P < 0.001$) and fear (1/10 correct; $z = 1.93$; $P < 0.05$), and borderline impairments of sadness (6/10 correct; $z = 1.66$; $P = 0.049$) and happiness ($z = 1.62$; $P = 0.053$). Note that the extent of the fear-recognition deficit may be underestimated in this test, owing to the fact that D.R. performed at floor level.

The second test assessed recognition of basic emotions conveyed by entirely non-verbal sound patterns (laughing for happiness, growling for anger, and so on). D.R. showed highly significantly impaired recognition of anger ($z = 3.73$, $P < 0.001$) and fear ($z = 6.89$, $P < 0.001$), and normal performance ($z < 1$, $P > 0.1$) for the other emotions.

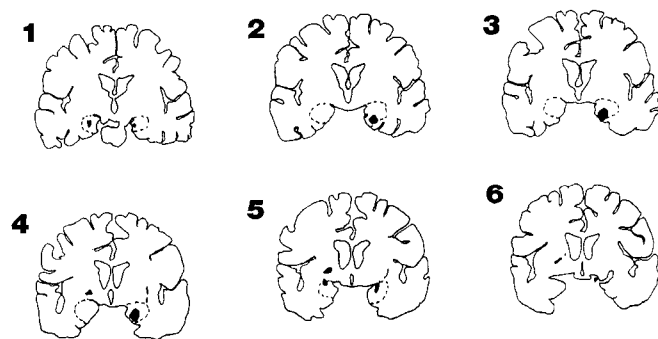


Figure 1 Tracings of lesions in the region of D.R.'s amygdala derived from 6 coronal MRI sections at 2.2-mm intervals. The region of the amygdala is marked by dashed lines, and the regions of damage are indicated in black. The sections are numbered in a caudal to rostral direction, with the left amygdala on the right side of each.

Taken together, our findings show that D.R. had impaired interpretation of paralinguistic signals involved in the communication of emotion, despite her normal hearing and speech perception. It is clear that these deficits are not entirely restricted to the recognition of emotion: D.R. had problems in using prosodic features to determine whether a sentence was intended as a question or an exclamation, and even to recognize or match a speaker's identity. When the deficit in recognizing emotion was investigated in more detail, however, it was the recognition of fear and anger that were consistently the most severely impaired. Note particularly that different cues are involved in signalling fear or anger by prosody or non-verbal sounds (screams and growls, for example), so D.R.'s consistent problems with these emotions cannot be reduced to some subtle hearing defect.

The pattern forms a close parallel to previous findings in the visual domain. For D.R., her interpretation is defective of social signals other than emotion from the face, including eye gaze direction³. Findings of impaired processing of social signals from face and voice confirm the suggestion that the amygdala forms part of the neural substrate for social cognition¹¹. In addition, the recognition of facial expressions of fear and anger is particularly severely compromised for D.R., and there is evidence that the same emotions are affected in the postencephalitic case S.E., who also has bilateral amygdala damage⁵. Although reports of case S.M., who has bilateral amygdala damage due to Urbach-Wiethe disease, emphasize her poor recognition of fear from the face, she is also significantly impaired at recognizing anger in some tests¹². Therefore, recognition of fear and anger in particular seems to be compromised by bilateral amygdala damage, regardless of the input modality.

The fact that D.R.'s recognition of emotion is impaired across the same basic emotions of fear and anger for voices and faces is consistent with impairment of a mechanism geared to interpreting emotional signals regardless of their source, rather than a mechanism specific to faces *per se*. Although it contains many cells responsive to faces seen²⁰⁻²³, the amygdala is known to receive inputs from other sense modalities²⁴ and is therefore in a position to be involved in multimodal recognition.

Fear is a socially contagious emotion²⁵ with a long evolutionary history²⁶. Displays of fear and anger by other people are signals of the presence of immediate threat in the environment, and anger carries a clear intention to frighten the recipient. A plausible hypothesis is therefore that impaired recognition of fear and anger after amygdala damage reflects involvement of the amygdala in the appraisal of danger and the emotion of fear^{12,13}. □

Methods

Audiometric testing used a DSP Pure Tone Audiometer, with separate testing of

Table 2 Performance of D.R. on auditory emotion recognition tasks

	D.R.	Controls	
		Mean	s.d.
Words			
Happiness	4/10	7.60	2.32
Sadness	6/10*	8.50	1.51
Anger	3/10***	8.20	1.14
Fear	1/10*	4.90	2.02
Disgust	6/10	5.90	1.97
Sounds			
Happiness	15/20	16.33	1.56
Sadness	15/20	16.00	2.00
Anger	5/20***	14.33	2.50
Fear	6/20***	16.33	1.50
Disgust	20/20	18.25	2.22
Surprise	18/20	17.58	1.88

Means and standard deviations are values for control subjects of comparable age and education. Asterisked scores are significantly below the control mean: * $z > 1.65$, $P < 0.05$; ** $z > 2.33$, $P < 0.01$; *** $z > 3.10$, $P < 0.001$.

left and right ears at 250, 500, 1,000, 2,000, 4,000 and 8,000 Hz. D.R.'s thresholds were all at normal levels for her age, with 100% correct reports of 20-db sounds across the frequencies critical for speech perception (500–2,000 Hz).

Auditory processing of words and non-words (Table 1) was assessed using tests from the Psycholinguistic Assessments of Language Processing in Aphasia battery¹⁴; control data are taken from the test's norms. The other tests summarized in Table 1 (recognition of environmental sounds, interpretation of vocal expressions of emotion, unfamiliar voice matching and recognition of familiar voices) are described elsewhere¹⁵. D.R.'s performance is compared with that of 20 control subjects (10 men, 10 women) aged 40–59 years¹⁵.

The emotional-words test (Table 2) involved 6 bisyllabic concrete nouns with neutral meaning (carpet, finger, hammock, motor, sailor, daughter). These nouns were selected to have varied phonetic properties. The initial phoneme was varied across two fricative onsets, two plosive, one aspirant and one nasal obstruent. Stress was held constant (on the first syllable) and each stressed syllable featured a different vowel.

These nouns were spoken by six native English speakers (3 male, 3 female, with a variety of accents), in five different manners intended to convey emotions of happiness, sadness, anger, fear or disgust. These emotions were chosen because they are each known to be associated with distinct sets of prosodic features¹⁹. Training and feedback were given to ensure that appropriate prosodic features were used to communicate each emotion. The speech was recorded in a sound-protected environment onto digital audio tape (DAT).

Tokens of each speech item were digitized at 22 kHz, and presented to 5 judges for categorization as each emotion. These categorizations were used to make up a set with ten different stimuli per emotion which resulted in similar average correct categorization of each emotion across the 5 judges (happiness, 82%; sadness, 93%; anger, 94%; fear, 89%; disgust, 84%). This selection was made subject to the constraint that at least one example of each word would be used for each emotion, and the different speakers should be used as evenly as possible (it was not possible to do this perfectly evenly across speakers: thus, there were 13 stimuli from each of three speakers, and 7 and 4 respectively from the other two).

For the test of emotion recognition from neutral-content words, the digitized tokens selected were presented in a random order using PsyScope software²⁷ running on a Macintosh Powerbook 540v computer and two loudspeakers set to a comfortable loudness level. Five additional items (one for each emotion) were used as practice trials. D.R.'s performance is compared to 10 controls aged 50–69 years (mean, 58.90 years; s.d., 6.59) without formal education.

Stimuli for the emotional-sounds test (Table 2) were recorded from two native English speakers (one male and one female), who were asked to generate a range of non-verbal sounds corresponding to the six basic emotions of happiness, sadness, anger, fear, disgust and surprise. This list was the same as that commonly used in work on facial expressions^{1,5,28,29} to facilitate comparison. Sounds included: for anger, growls; fear, screams; happiness, laughter; sadness, sobbing; surprise, gasping; disgust, retching. The vocal expressions were recorded onto DAT in a sound-protected environment.

Tokens of the different vocal expressions were digitized at 22 kHz. In all, there were 27 angry sounds, 20 disgust, 24 fear, 27 happy and 31 sad sounds. These were presented to 6 judges for categorization. From their classifications, ten stimuli per emotion were selected so that the six emotions were recognized at similar overall levels (happiness, 90%; sadness, 87%; anger, 95%; fear, 93%; disgust, 97%; surprise, 93%).

For the test of emotion-recognition from sounds, the digitized tokens selected were presented in a random order using PsyScope software²⁷ running on a Macintosh Powerbook 540v computer and two loudspeakers set to a comfortable loudness level. Six additional items (one for each emotion) were used as practice trials, following which the test sounds were presented in a random order, with each stimulus presented twice across two blocks of trials. D.R.'s performance is compared to 12 controls aged 50–64 years (mean, 57.17; s.d., 3.88) without formal education.

Received 8 October; accepted 21 November 1996.

- Adolphs, R., Tranel, D., Damasio, H. & Damasio, A. *Nature* **372**, 669–672 (1994).
- Adolphs, R., Tranel, D., Damasio, H. & Damasio, A. R. *J. Neurosci.* **15**, 5879–5891 (1995).
- Young, A. W. et al. *Brain* **118**, 15–24 (1995).
- Young, A. W., Hellawell, D. J., van de Wal, C. & Johnson, M. *Neuropsychologia* **34**, 31–39 (1996).
- Calder, A. J. et al. *Cogn. Neuropsychol.* **13**, 699–745 (1996).
- Hamann, S. B. et al. *Nature* **379**, 497 (1996).
- Bechara, A. et al. *Science* **269**, 1115–1118 (1995).
- Cahill, L., Babinsky, R., Markowitsch, H. J. & McGaugh, J. L. *Nature* **377**, 295–296 (1995).
- Cahill, L. et al. *Proc. Natl Acad. Sci. USA* **93**, 8016–8021 (1996).
- Morris, J. S. et al. *Nature* **383**, 812–815 (1996).
- Allman, J. & Brothers, L. *Nature* **372**, 613–614 (1994).
- Davis, M. *Annu. Rev. Neurosci.* **15**, 353–375 (1992).
- LeDoux, J. E. *Annu. Rev. Psychol.* **46**, 209–235 (1995).
- Kay, J., Lesser, R. & Coltheart, M. *Psycholinguistic Assessments of Language Processing in Aphasia (PALPA)* (Erlbaum, Hove, East Sussex, 1992).
- Reid, I., Young, A. W. & Hellawell, D. J. *Behav. Neurol.* **6**, 225–228 (1993).
- Frick, R. W. *Psychol. Bull.* **97**, 412–429 (1985).
- Van Lancker, D., Kreiman, J. & Wickens, T. D. *J. Phonet.* **13**, 39–52 (1985).
- Nolan, F. *The Phonetic Bases of Speaker Recognition* (Cambridge University Press, Cambridge, 1983).
- Murray, I. R. & Arnott, J. L. *J. Acous. Soc. Am.* **93**, 1097–1108 (1993).
- Leonard, C. M., Rolls, E. T., Wilson, F. A. W. & Baylis, C. G. *Behav. Brain Res.* **15**, 159–176 (1985).
- Rolls, E. T. *Hum. Neurobiol.* **3**, 209–222 (1984).
- Nakamura, K., Mikami, A. & Kubota, K. *J. Neurophysiol.* **67**, 1447–1463 (1992).
- Seeck, M. et al. *Ann. Neurol.* **34**, 369–372 (1993).
- Aggleton, J. P. *The Amygdala* (Wiley-Liss, New York and Chichester, 1992).
- Brothers, L. *Am. J. Psychiat.* **146**, 10–19 (1989).
- Ekman, P. *Darwin and Facial Expression: A Century of Research in Review* (Academic, New York, 1973).
- Cohen, J. D., MacWhinney, B., Flatt, M. & Provost, J. *Behav. Res. Meth.* **25**, 257–271 (1993).
- Ekman, P. & Friesen, W. V. *Pictures of Facial Affect* (Consulting Psychologists Press, Palo Alto, California, 1976).
- Ekman, P. in *Nebraska Symposium on Motivation, 1971* (ed. Cole, J. K.) 207–283 (University of Nebraska Press, Lincoln, Nebraska, 1972).

Acknowledgements: We thank D.R. and her husband for their time and patience in cooperating with our investigation. This research was supported by a grant from the EJLB foundation.

Correspondence and requests for materials should be addressed to A.W.Y. (e-mail: andy.young@mrc-apu.cam.ac.uk).

Independent requirement for ISL1 in formation of pancreatic mesenchyme and islet cells

Ulf Ahlgren*, Samuel L. Pfaff†‡, Thomas M. Jessell†, Thomas Edlund* & Helena Edlund*

* Department of Microbiology, University of Umeå, S-901 87 Umeå, Sweden

† Howard Hughes Medical Institute, Department of Biochemistry and Molecular Biophysics, Center for Neurobiology and Behavior, Columbia University, New York, New York 10032, USA

The mammalian pancreas is a specialized derivative of the primitive gut endoderm and controls many homeostatic functions through the activity of its component exocrine acinar and endocrine islet cells. The LIM homeodomain protein ISL1 is expressed

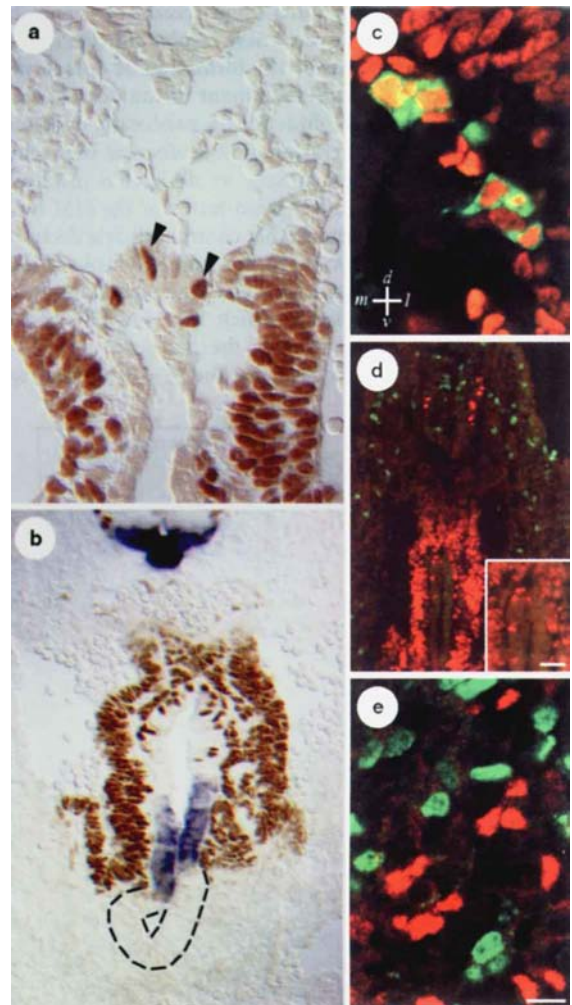


Figure 1 Expression of ISL1 in pancreatic epithelium and mesenchyme. **a**, In 16-somite embryos, ISL1⁺ cells are detected in the dorsal pancreatic epithelium (arrowheads) and in the lateral mesenchyme. **b**, At the 25-somite stage, ISL1⁺ mesenchyme has accumulated around the dorsal pancreatic bud. ISL1 is not expressed in the ventral pancreatic epithelium (broken line) or in the adjacent mesenchyme. *Shh* mRNA expression (dark blue) is used to visualize lateral endoderm¹⁶. **c**, Confocal image of the dorsal pancreatic bud of an e9.5 *Isl1*(+/+) embryo showing that all glucagon⁺ cells (green) are also ISL1⁺ (red). The image is representative of 5 embryos examined. **d**, No pancreatic epithelial cells in e9.5 *Isl1*(+/+) embryos are double-labelled with mpm2 (ref. 22) (green) and ISL1 (red) antibodies (insert) (*n* = 653 ISL1⁺ cells analysed; 4 embryos). **e**, No ISL1⁺ cells (red) incorporated BrdU (green) in explants of e12 dorsal pancreas cultured for 24 h and labelled with BrdU for 60 min to detect cells in the S and G2 phases of the cell cycle (*n* = 814 ISL1⁺ cells; *n* = 1694 BrdU⁺ cells analysed; 2 embryos). In **a–d**, dorsal epithelium is up. Scale bar in **a**, **d**, insert, 20 µm; in **b**, 40 µm; in **c**, **e**, 10 µm.

in all classes of islet cells in the adult^{1,2} and its expression in the embryo is initiated soon after the islet cells have left the cell cycle. ISL1 is also expressed in mesenchymal cells that surround the dorsal but not ventral evagination of the gut endoderm, which together comprise the pancreatic anlagen. To define the role of ISL1 in the development of the pancreas, we have now analysed acinar and islet cell differentiation in mice deficient in ISL1 function³. Dorsal pancreatic mesenchyme does not form in ISL1-mutant embryos and there is an associated failure of exocrine cell differentiation in the dorsal but not the ventral pancreas. There is also a complete loss of differentiated islet cells. Exocrine, but not endocrine, cell differentiation in the dorsal

‡ Present address: Gene Expression Laboratory, The Salk Institute, 10010 North Torrey Pines Road, La Jolla, California 92037, USA

pancreas can be rescued *in vitro* by provision of mesenchyme derived from wild-type embryos. These results indicate that ISL1, by virtue of its requirement for the formation of dorsal mesenchyme, is necessary for the development of the dorsal exocrine pancreas, and also that ISL1 function in pancreatic endodermal cells is required for the generation of all endocrine islet cells.

The *Isl1* gene was defined initially in studies on insulin gene regulation and encodes a transcription factor of the LIM homeo-domain family, a class of proteins that control cell-fate decisions in invertebrates and vertebrates^{3–10}. To determine the role of ISL1 in the differentiation of pancreatic islet cells, we analysed the development of the pancreas in mice in which ISL1 function had been eliminated by targeted disruption of the *Isl1* gene³. In wild-type mice, ISL1⁺ cells can first be detected at the 15–16-somite stage, at

embryonic day (e) 9, in the dorsal pancreatic epithelium and in cells of the lateral mesenchyme surrounding the gut epithelium (Fig. 1a). At the 20–25 somite stage, ISL1 expression is detected in the mesenchymal cells that have now enveloped the dorsal pancreatic epithelium (Fig. 1b). At no stage do mesenchymal cells flanking the ventral pancreatic epithelium express ISL1 (Fig. 1b), indicating that the mesenchyme surrounding the early pancreatic epithelium is patterned along the dorsal–ventral axis.

Glucagon-positive cells appear first at the 20–25-somite stage (e9.5) and they all express ISL1 (Fig. 1c), whereas cells expressing insulin, somatostatin and pancreatic polypeptide differentiate at later stages^{11–14} and also express ISL1 when they appear (ref. 2, and data not shown). ISL1 expression in the ventral pancreatic epithelium can first be detected at embryonic day 11 (data not shown), consistent with the 1–2-day delay in the genesis of islet cells in the ventral pancreatic bud^{12,15}. ISL1⁺ endocrine cells analysed between e9.5–e13 did not express the M-phase-specific antigen detected by the anti-mpm2 antibody and did not incorporate BrdU (Fig. 1d, e). Together, these results provide evidence that ISL1 expression in islet cells is initiated after their final mitotic division but before the onset of hormone-gene expression.

Mice lacking ISL1 are arrested in development soon after e9.5 (ref. 3) and we therefore analysed cell differentiation in e9.5 *Isl1*(+/+) and *Isl1*(–/–) embryos using PDX1 expression to delineate specialized pancreatic epithelium and *Shh* expression as a more general marker of gut epithelium¹⁶. In *Isl1*(–/–) embryos, the pancreatic epithelium and its associated ventral and lateral mesenchyme appeared normal histologically, PDX1 expression in the ventral epithelium was normal and the lateral gut epithelium still expressed *Shh* (Fig. 2a, b). In contrast, there was a virtually complete depletion of the mesenchyme flanking the dorsal pancreatic epithelium (Fig. 2b) and PDX1 was expressed at markedly lower levels in the dorsal epithelium in *Isl1*(–/–) embryos than in *Isl1*(+/+) embryos (Fig. 2a, b). To determine whether ISL1 has an independent function in the generation of islet cells, we monitored the expression of glucagon in *Isl1*(+/+) and *Isl1*(–/–) embryos. Glucagon-positive cells were detected in *Isl1*(+/+) but not in *Isl1*(–/–) e9.5 embryos (Fig. 2c, d). Thus, as summarized in Fig. 2e, ISL1 is necessary for the formation of the dorsal pancreatic mesenchyme and for the generation of glucagon-expressing cells in the dorsal pancreatic epithelium.

To determine whether the absence of glucagon-positive cells in e9.5 *Isl1*(–/–) embryos reflects a delay or a complete block in their differentiation, we isolated segments of gut containing the pancreatic anlagen from e9.5 *Isl1*(+/+) and *Isl1*(–/–) embryos and maintained them in culture for 7 days, which enabled us to analyse whether the differentiation of the later-appearing insulin-, and somatostatin-expressing cells is also impaired in *Isl1*(–/–) embryos. Explants derived from *Isl1*(+/+) embryos gave rise to endocrine cells positive for glucagon, insulin and somatostatin (Fig. 3a–c), and exocrine cells positive for amylase (Fig. 3d). Explants derived from *Isl1*(–/–) embryos, however, gave rise to cells that were negative for glucagon, insulin and somatostatin (Fig. 3e–g), although exocrine cells were present that were positive for amylase (Fig. 3h). These results indicate that ISL1 expression in pancreatic epithelium is required for the differentiation of islet but not exocrine cells.

The differentiation of pancreatic epithelium is thought to depend on signals provided by the surrounding mesenchyme¹⁷. In *Isl1*(–/–) embryos pancreatic mesenchyme surrounds only the ventral pancreatic epithelium, so the amylase-positive exocrine cells generated in gut explants from *Isl1*(–/–) embryos may have derived solely from the ventral pancreatic epithelium. In support of this, amylase-positive exocrine cells were generated in cultured ventral half-gut explants but not in dorsal half-gut explants isolated from *Isl1*(–/–) embryos (Fig. 4a, b). Both gut explant halves isolated from *Isl1*(+/+) embryos generated amylase-positive cells

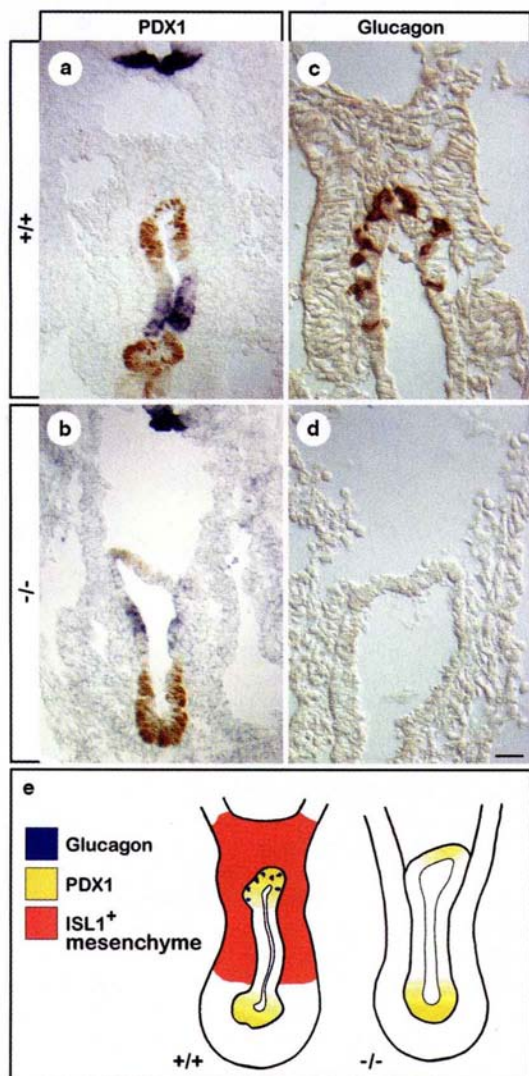


Figure 2 Loss of dorsal pancreatic mesenchyme and glucagon⁺ cells in *Isl1*(–/–). Photomicrographs of transverse sections of *Isl1*(+/+) (a, c) and *Isl1*(–/–) (b, d) at the level of the pancreas. In a and b, *in situ* hybridization (dark blue) using a *Shh* probe was done to visualize the lateral endoderm¹⁶. a, Transverse section of an e9.5 *Isl1*(+/+) embryo stained with PDX1 antibodies to indicate the dorsal and ventral pancreatic buds. b, Transverse section of an e9.5 *Isl1*(–/–) embryo labelled with PDX1 antibodies. The dorsal pancreatic mesenchyme is absent, whereas the ventral pancreatic epithelium with its associated mesenchyme appears normal. c, At the 25-somite stage, glucagon⁺ cells have appeared in the dorsal pancreatic epithelium in *Isl1*(+/+) embryos, whereas in somite-matched *Isl1*(–/–) embryos (d), no glucagon⁺ cells are detected. e, Comparison of e9.5 wild-type and ISL1 mutant phenotype.

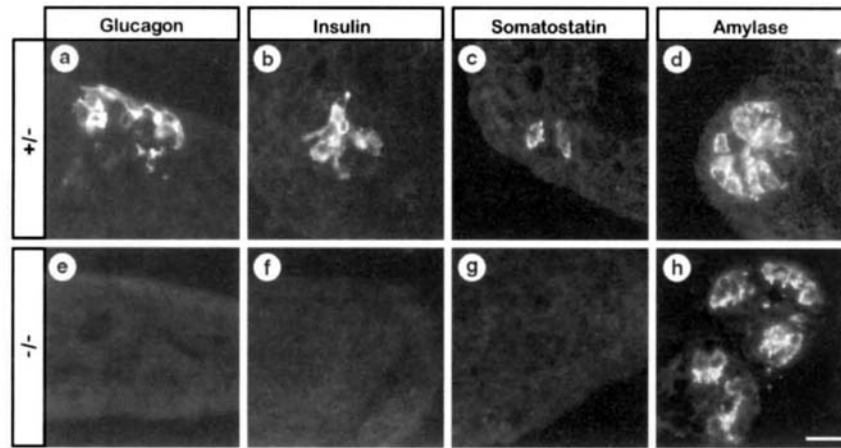


Figure 3 Differentiated islet cells are not generated in explants derived from *Isl1*^{-/-} embryos. Fluorescence immunohistochemistry of alimentary tract explants derived from *Isl1*^{+/+} (a-d) and *Isl1*^{-/-} (e-h) mice. Islet cell hormones, glucagon (a), insulin (b) and somatostatin (c) and the exocrine

enzyme amylase (d) are detected in explants derived from *Isl1*^{+/+} embryos. In explants derived from *Isl1*^{-/-} embryos there is no expression of glucagon (e), insulin (f) and somatostatin (g), whereas amylase⁺ cells (h) are still present. Images are representative of at least 4 explants. Scale bar, 20 μ m.

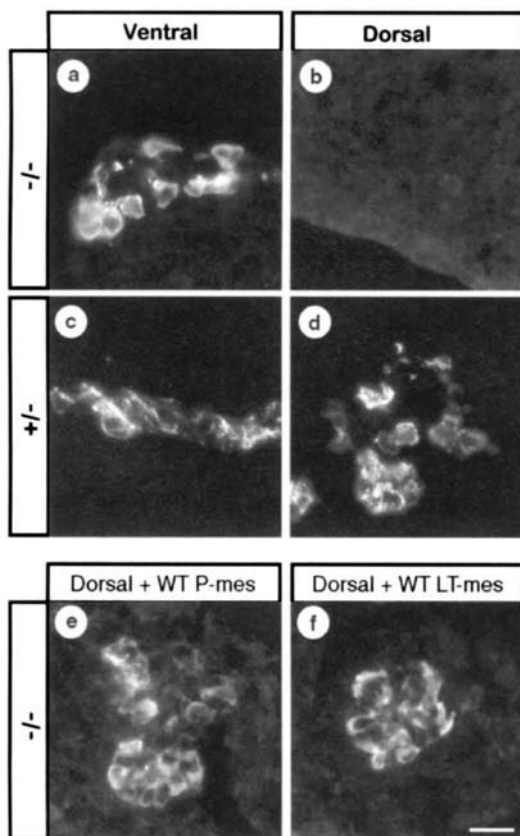


Figure 4 Dorsal pancreatic exocrine differentiation is impaired in explants derived from *Isl1*^{-/-} embryos. Immunohistochemical analysis of marker expression of explants derived from the dorsal (b, d, e, f) or ventral (a, c) pancreatic anlagen. Amylase⁺ cells are generated in ventral (a) but not dorsal (b) explants derived from *Isl1*^{-/-} embryos. In contrast, amylase⁺ cells are detected in both ventral (c) and dorsal (d) explants from *Isl1*^{+/+} embryos. Recombination of dorsal pancreatic anlagen from *Isl1*^{-/-} (e, f) with dorsal pancreatic (e) or lung-tracheal region (f) mesenchyme from somite-matched (+/+) type embryos rescues the differentiation of dorsal exocrine cells. Images are representative of 5 (c, d), 8 (a, b) and 3 (e, f) explants. Abbreviations: P, pancreas; LT, lung-tracheal; mes, mesenchyme. Scale bar, 20 μ m.

(Fig. 4c, d). To determine whether the loss of amylase-positive cells in the dorsal bud of *Isl1*^{-/-} embryos reflected the absence of mesenchyme, we recombined the dorsal half gut of *Isl1*^{-/-} embryos with dorsal pancreatic mesenchyme derived from wild-type e9.5 embryos. Amylase-positive cells were generated in such recombinants (Fig. 4e) and also when the dorsal half gut of *Isl1*^{-/-} embryos was recombined with mesenchyme from the lung-tracheal region (Fig. 4f). The presence of wild-type mesenchyme did not restore the differentiation of endocrine islet cells in cultures of the dorsal half-gut region (data not shown).

These findings suggest that ISL1 functions at two independent steps in the development of the pancreas. First, ISL1 is expressed in, and is required for, the development of dorsal pancreatic mesenchyme, providing genetic evidence that the elimination of pancreatic mesenchyme blocks exocrine pancreas development. Second, ISL1 expression in pancreatic epithelial cells is required for the differentiation of islet cells, suggesting a cell-autonomous function similar to that reported in motor neuron generation³. The development of the pancreas seems therefore to be controlled through the sequential activities of distinct classes of transcription factors. The homeodomain transcription factor PDX1 seems to specify the early pancreatic epithelium, permitting its proliferation, branching and subsequent differentiation^{16,18,19}. The basic helix-loop-helix protein *NeuroD/BETA2* (refs 20, 21) is expressed in pancreatic endocrine cells (refs 20, 21, and our own unpublished observations) and, by analogy with its neurogenic function²⁰, may act upstream of *Isl1* in the islet-cell differentiation pathway. More generally, the demonstration of a requirement for ISL1 in both islet-cell and motor-neuron development suggests that there are conserved elements in the transcriptional control of neuronal and pancreatic endocrine differentiation. □

Methods

Animals. The generation of *Isl1*^{-/-} mice has been described elsewhere³; +/+, +/- and -/- mice were obtained from our local breeding colony of *Isl1*^{+/+} mice. Mice and embryos were genotyped as described³.

Immunohistochemistry. Primary antibodies, rabbit anti-ISL1², rabbit anti-PDX1¹⁶ and guinea-pig anti-glucagon (Linco) were detected using the ABC immunoperoxidase system (Vector Labs). Guinea-pig anti-rat insulin-C peptide (Linco), guinea-pig anti-glucagon (Linco), rabbit anti-somatostatin (DAKO) and rabbit anti-human α -amylase (Sigma) was detected as described¹⁶. *In situ* hybridization using an *Isl2* probe³ showed that no cells in the pancreatic epithelium express *Isl2* (data not shown). In addition, cells in the

pancreatic epithelium and in the dorsal pancreatic mesenchyme that are recognized by anti-ISL1/ISL2 ployclonal antibody are also recognized by an ISL1-specific mouse monoclonal anti-ISL1 antibody³ (data not shown), indicating that these cells express ISL1 but not ISL2 and can therefore be referred to as ISL1⁺ cells. Double-label immunohistochemistry, *in situ* hybridization and Confocal microscopy analysis were done as described¹⁶, as were ISL1/mpm2 and ISL1/BrdU double labelling⁷. Primary antibodies were mouse anti-mpm2 (Upstate Biotech), mouse anti-BrdU (Becton Dickinson), and rabbit anti-ISL1. When analysing transverse sections of e9.5 embryos, PDX1 was used on consecutive sections as a pancreatic marker.

Culture of pancreatic rudiments. Gut explants containing the dorsal and/or ventral pancreatic anlagen from e9.5 *Isl1*(+/−) and *Isl1*(−/−) embryos were isolated and cultured for 7 days¹⁶. Explants were then processed, sectioned and immunostained using standard procedures^{16,18}.

Received 22 October; accepted 13 November 1996.

1. Karlsson, O., Thor, S., Norberg, T., Ohlsson, H. & Edlund, T. *Nature* **344**, 879–882 (1990).
2. Thor, S., Ericson, J., Brännström, T. & Edlund, T. *Neuron* **7**, 881–889 (1991).
3. Pfaff, S. L., Mendelsohn, M., Stewart, C. L., Edlund, T. & Jessell, T. M. *Cell* **84**, 309–320 (1996).
4. Way, J. C. & Chalfie, M. *Cell* **54**, 5–16 (1988).
5. Frey, D., Kim, S. K. & Horvitz, H. R. *Nature* **344**, 876–879 (1990).
6. Blair, K. A. *Bioessays* **17**, 299–309 (1995).
7. Lundgren, F. E., Callahan, C. A., Thor, S. & Thomas, J. D. *Development* **121**, 1769–1773 (1995).
8. Li, H. *et al.* *EMBO J.* **13**, 2876–2885 (1994).
9. Shawlot, W. & Behringer, R. R. *Nature* **374**, 425–430 (1995).
10. Sheng, H. *et al.* *Science* **272**, 1004–1007 (1996).
11. Alpert, S., Hanahan, D. & Teitelman, G. *Cell* **53**, 295–308 (1988).
12. Herrera, P. L. *et al.* *Development* **113**, 1257–1265 (1991).
13. Teitelman, G., Alpert, S., Polack, J. M., Martinez, A. & Hanahan, D. *Development* **118**, 1031–1039 (1993).
14. Upchurch, B. H., Aponte, G. W. & Litter, A. B. *Development* **120**, 245–252 (1994).
15. Pictet, R. & Rutter, W. J. in *Handbook of Physiology* Vol. 1, 25–66 (American Physiological Society, Washington DC, 1972).
16. Ahlgren, U., Jonsson, J. & Edlund, H. *Development* **122**, 1409–1416 (1996).
17. Golosow, N. & Grobstein, C. *Dev. Biol.* **4**, 242–255 (1962).
18. Jonsson, J., Carlsson, L., Edlund, T. & Edlund, H. *Nature* **371**, 606–609 (1994).
19. Offield, M. F. *et al.* *Development* **122**, 983–995 (1996).
20. Lee, J. E. *et al.* *Science* **268**, 836–844 (1995).
21. Naya, F. J., Stellrecht, C. M. M. & Tsai, M.-J. *Genes Dev.* **9**, 1009–1019 (1995).
22. Kuang, J., Zhao, J., Wright, D. A., Saunders, G. F. & Rao, P. N. *Proc. Natl Acad. Sci. USA* **86**, 4982–4986 (1994).

Acknowledgements: We thank K. Simu, K. Falk, U. B. Backman, B. E. Nordlund and J. Pettersson for technical assistance, U. Dahl, J. Jonsson and members of our laboratory for discussion, S. Sandström for artwork and M. Mendelsohn for generating *Isl1*(−/−) mice. This work was supported by grants from the Swedish Medical Research Council (to H.E. and T.E.) and by the Juvenile Diabetes Foundation, New York (to H.E.). T.M.J. is an Investigator of the Howard Hughes Medical Institute.

Correspondence and requests for materials should be addressed to H.E. (e-mail: Helena. Edlund@micro.umu.se).

Distinct functions of nuclear and cytoplasmic calcium in the control of gene expression

Giles E. Hardingham, Sangeeta Chawla, Claire M. Johnson & Hilmar Bading

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

Calcium entry into neuronal cells through voltage or ligand-gated ion channels triggers neuronal activity-dependent gene expression critical for adaptive changes in the nervous system^{1–5}. Cytoplasmic calcium transients are often accompanied by an increase in the concentration of nuclear calcium^{6–9}, but the functional significance of such spatially distinct calcium signals is unknown. Here we show that gene expression is differentially controlled by nuclear and cytoplasmic calcium signals which enable a single second messenger to generate diverse transcriptional responses. We used nuclear microinjection of a non-diffusible calcium chelator to block increases in nuclear, but not cytoplasmic, calcium concentrations following activation of L-type voltage-gated calcium channels. We showed that increases in nuclear calcium concentration control calcium-activated gene

expression mediated by the cyclic-AMP-response element (CRE), and demonstrated that the CRE-binding protein CREB can function as a nuclear calcium-responsive transcription factor. A second signalling pathway, activating transcription through the serum-response element (SRE), is triggered by a rise in cytoplasmic calcium and does not require an increase in nuclear calcium.

Confocal laser scanning microscopy of the Ca²⁺ indicator Fluo-3 was used to measure changes in cytoplasmic calcium ([Ca²⁺]_c) and nuclear calcium ([Ca²⁺]_n) concentrations in AtT20 cells, a mouse pituitary cell line. Ca²⁺ influx through L-type calcium channels was triggered by KCl-induced membrane depolarization in the presence of the L-type calcium-channel agonist FPL 64176 (ref. 10) (KCl/FPL treatment). Upon KCl/FPL treatment, [Ca²⁺]_c increased from resting levels of 108 ± 5 nM to 783 ± 40 nM and [Ca²⁺]_n rose from 104 ± 6 nM to 979 ± 58 nM (*n* = 47). To block increases in [Ca²⁺]_n specifically without altering the rise in [Ca²⁺]_c, we microinjected the Ca²⁺ chelator BAPTA into the nucleus. The BAPTA used was coupled to a 70K dextran molecule (BAPTA-D70), which prevented it from diffusing out of the nucleus into the cytoplasm. This was confirmed by microinjecting BAPTA-D70 together with a Texas red-labelled 70K dextran into the nucleus. The localization of the red dye, which reflects the subcellular distribution of the 70K dextran, was monitored and found to remain exclusively nuclear for up to 16 hours after injection (data not shown).

Nuclear microinjection of BAPTA-D70 specifically reduced KCl/FPL-induced increases in [Ca²⁺]_n by ~50% (558 ± 76 nM; *n* = 8) but did not affect the rise in [Ca²⁺]_c (864 ± 104 nM; *n* = 8) (Fig. 1a, b). In cells injected with the control solution, [Ca²⁺]_c and [Ca²⁺]_n increased from resting levels of 106 ± 17 nM and 79 ± 12 nM to 879 ± 93 nM and 1,105 ± 149 nM (*n* = 12), respectively, upon KCl/FPL stimulation (Fig. 1a, b). These Ca²⁺ levels are similar to those observed in uninjected cells, demonstrating that the injection process *per se* has no effect on Ca²⁺ levels. A time course of the [Ca²⁺]_n: [Ca²⁺]_c (N/C) ratio, an indicator of the specificity of the effect of BAPTA-D70 on [Ca²⁺]_n, demonstrates that the reduction of [Ca²⁺]_n in BAPTA-D70-injected, KCl/FPL-stimulated cells persisted for at least 30 min (Fig. 1c).

Having established a technique to reduce specifically increases in [Ca²⁺]_n without interfering with [Ca²⁺]_c, we next determined whether nuclear Ca²⁺ is critical for the control of Ca²⁺-activated transcription. The *c-fos* gene, a well characterized Ca²⁺-inducible gene^{1–3}, was used as a model system. Nuclear microinjection of BAPTA-D70, but not of the control, reduced KCl/FPL-induced expression of the endogenous *c-fos* protein, detected by immunofluorescence and quantified using a confocal laser scanning microscope, to 55 ± 7% (136 cells analysed). In contrast, cAMP-induced endogenous *c-fos* protein expression in BAPTA-D70-injected cells was 113 ± 7% compared with that of control-injected cells (57 cells analysed), demonstrating the specificity of the inhibition by nuclear BAPTA-D70 of Ca²⁺-dependent transcriptional activation. Because Ca²⁺-activated expression of *c-fos* was only partially sensitive to inhibition of increases in [Ca²⁺]_n, we considered that Ca²⁺ might control gene expression through several mechanisms, differing in their requirements for nuclear Ca²⁺. Expression of *c-fos* by Ca²⁺ signals is regulated by two DNA regulatory elements, the CRE and the SRE¹¹. To investigate whether the mechanisms that control CRE- and SRE-dependent Ca²⁺-activated transcription differ in their dependency for increases in [Ca²⁺]_n, we introduced into the cells plasmids containing the wild-type human *c-fos* gene (including either 711 base pairs (bp) in plasmid pF711myc, or 222 bp of upstream regulatory sequence in plasmid pF222myc), or the human *c-fos* gene in which the SRE or the CRE had been mutated in the context of an otherwise intact promoter. An oligonucleotide encoding a Myc-epitope tag was inserted in-frame into the *c-fos* coding region, enabling the expression of these *c-fos* gene constructs to be detected by immunofluorescence using an antibody against the Myc epitope (9E10) (Fig. 2a).

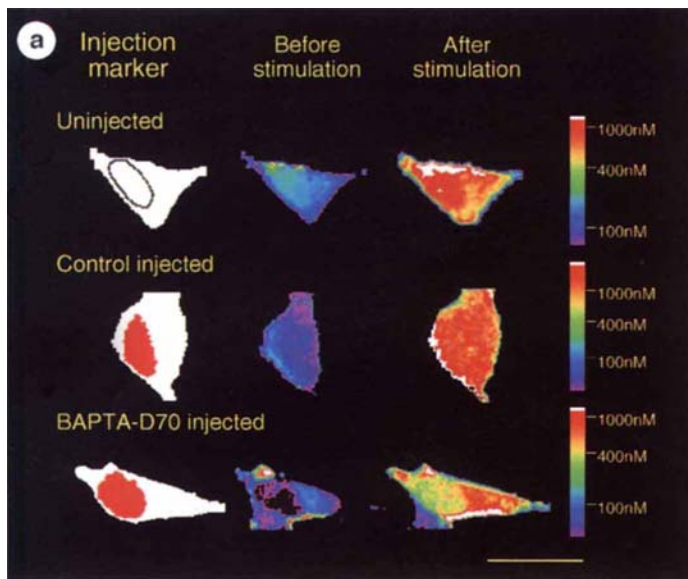
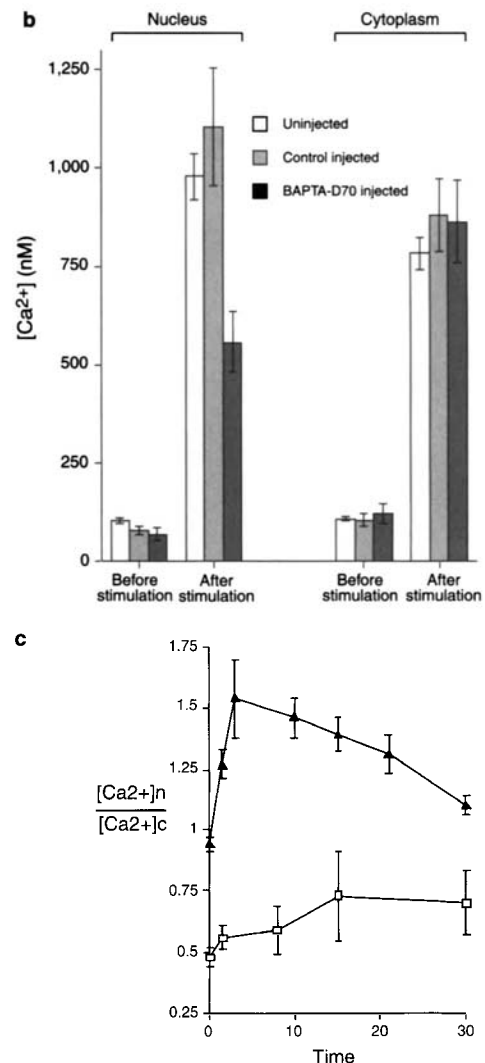


Figure 1 Nuclear microinjection of BAPTA-D70 specifically reduces rises in nuclear Ca^{2+} following activation of L-type calcium channels. **a**, Examples of Ca^{2+} -imaging data from uninjected, control-injected and BAPTA-D70-injected AtT20 cells. The Fluo-3 fluorescence images are taken before and 90 s after stimulation and are shown as a fraction of the image taken following the saturation of Fluo-3 with Ca^{2+} by the addition of $50 \mu\text{M}$ ionomycin. The nuclear localization of the injection marker is shown by a Texas red fluorescence image. Scale bar, $25 \mu\text{m}$. **b**, Summary of Ca^{2+} imaging experiments. Nuclear microinjection of BAPTA-D70 ($n = 8$) reduced increases in $[\text{Ca}^{2+}]_n$ but not in $[\text{Ca}^{2+}]_c$ following KCl/FPL treatment, compared to control-injected ($n = 12$) and uninjected cells ($n = 47$) ($P < 0.01$, two-tailed unpaired t -test; $P < 0.01$, Mann-Whitney U test). Resting $[\text{Ca}^{2+}]_n$ and $[\text{Ca}^{2+}]_c$ are not affected by nuclear microinjection of BAPTA-D70 (two-tailed, unpaired t -test and Mann-Whitney U test). **c**, Time course of $[\text{Ca}^{2+}]_n/[\text{Ca}^{2+}]_c$ (N/C) ratio following KCl/FPL-stimulation of uninjected cells (triangles; $n = 57$ (0 and 90 s time points); $n = 10$ (3, 10, 15, 20, 30 min)) and cells microinjected into the nucleus with BAPTA-D70 (squares; $n = 13$ (0 and 90 s); $n = 5$ (8, 15, 30 min)). The N/C ratio in control-injected cells is similar to that in uninjected cells.

These constructs were first transfected into AtT20 cells and tested for inducibility following KCl/FPL treatment in RNase protection assays (Fig. 2b).

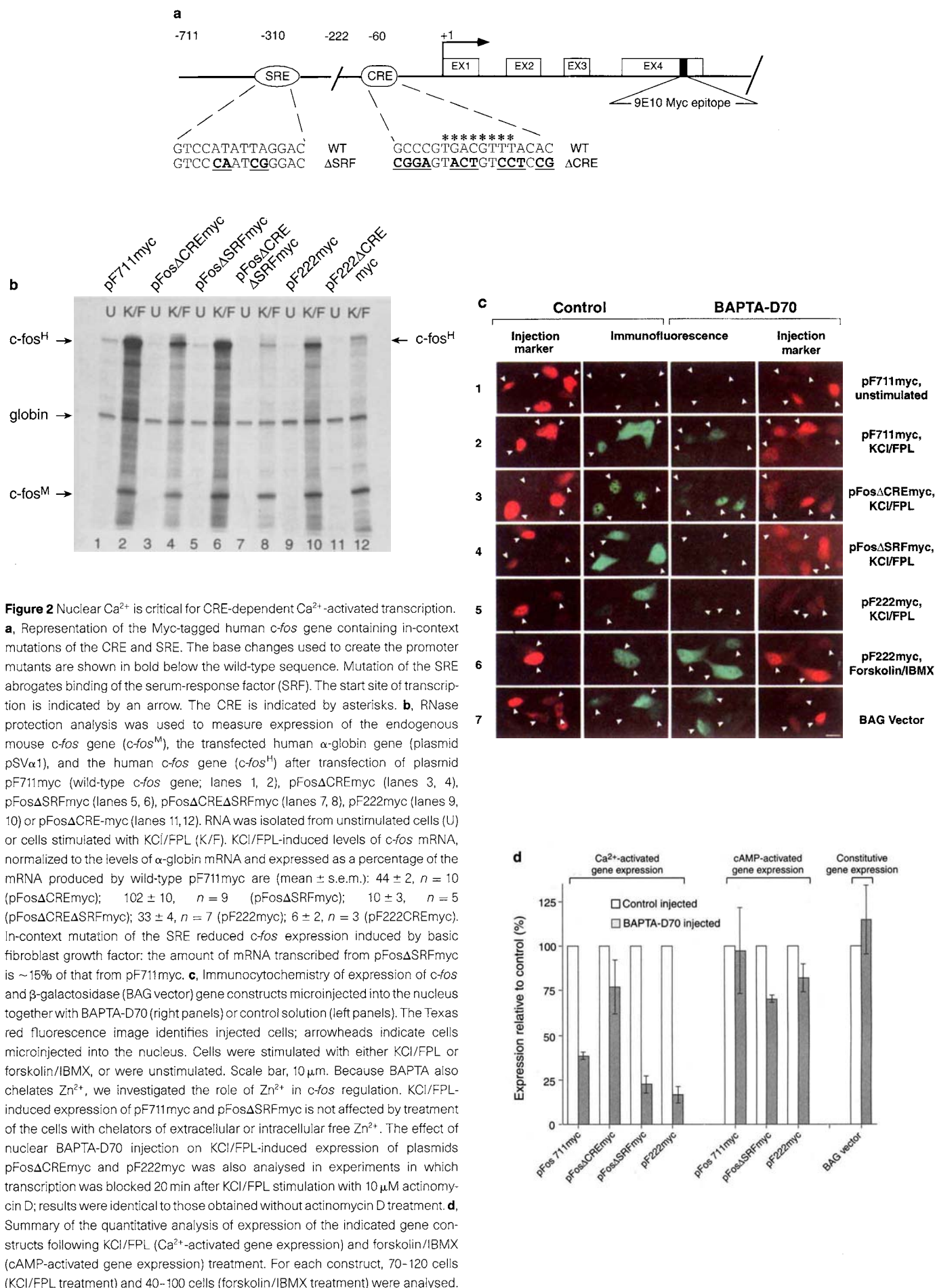
The basal level of expression of the transfected wild-type *c-fos* gene (pF711myc) and of the endogenous *c-fos* gene, which was analysed in parallel, was low, and KCl/FPL treatment of the cells resulted in a strong induction of transcription (Fig. 2b, lanes 1, 2). In-context mutation of the CRE (construct pFos Δ CREmyc) decreased the response by $\sim 50\%$ compared to wild type (Fig. 2b; compare lanes 2 and 4). The levels of mRNA obtained upon KCl/FPL stimulation of cells transfected with plasmid pFos Δ CRE Δ SRFmyc, which contains in-context mutations of the SRE and CRE, were very low (Fig. 2b; compare lanes 4 and 8). This shows that the transcriptional response of pFos Δ CREmyc is dependent on the SRE and that expression of the pFos Δ SRFmyc construct, which contains an SRE in-context mutation and that gives rise to mRNA levels similar to the wild type (Fig. 2b; compare lanes 2 and 6), is dependent on the CRE. The transcriptional response of construct pF222myc, containing only 222 bp of the upstream regulatory region and lacking the SRE, is also CRE-dependent, as mutation of the CRE (pF222 Δ CREmyc) reduced the amount of mRNA produced to negligible levels (Fig. 2b; compare lanes 10 and 12).

Having characterized two CRE-dependent constructs (pFos Δ SRFmyc and pF222myc) and an SRE-dependent construct (pFos Δ CREmyc), we next used microinjection experiments to assess the importance of $[\text{Ca}^{2+}]_n$ in transcriptional activation



through these regulatory elements. For each construct, 70–120 cells (Ca^{2+} -activated gene expression) and 40–100 cells (cAMP-activated gene expression) were analysed. The transcriptional response of the CRE-dependent *c-fos* gene constructs, pFos Δ SRFmyc and pF222myc, is dramatically reduced by nuclear microinjection of BAPTA-D70, as compared to injection of control (Fig. 2c; panels 4 and 5; Fig. 2d). In contrast, there was little effect on induction of the SRE-dependent *c-fos* gene construct, pFos Δ CREmyc, by nuclear microinjection of BAPTA-D70 (Fig. 2c, panel 3; Fig. 2d). There was an intermediate effect of nuclear BAPTA-D70 on the inducibility of the wild-type gene construct pF711myc (containing both the CRE and the SRE) (Fig. 2c, panel 2; Fig. 2d), which is similar to the results obtained with the endogenous *c-fos* gene. All CRE-containing *c-fos* constructs were tested for inducibility by cAMP, which was little affected by nuclear BAPTA-D70 microinjection (Fig. 2c, panel 6; Fig. 2d). In addition, nuclear microinjection of the BAPTA-D70 did not alter the expression of a β -galactosidase gene under the control of the Moloney murine leukaemia virus long terminal repeat (LTR), a constitutively active promoter (Fig. 2c, panel 7; Fig. 2d).

These results indicate that increases in $[\text{Ca}^{2+}]_n$ are critical for CRE-dependent but not necessary for SRE-dependent Ca^{2+} -activated transcription. However, because nuclear BAPTA-D70 microinjection reduced KCl/FPL-induced increases in $[\text{Ca}^{2+}]_n$ by $\sim 50\%$ (Fig. 1), SRE-dependent responses may also be stimulated by increases in $[\text{Ca}^{2+}]_n$ but have a lower activation threshold. We



therefore established two further stimulation paradigms, treatment with KCl only or with FPL 64176 only, which, compared to KCl/FPL treatment, gives rise to smaller increases in intracellular Ca^{2+} concentration ($[\text{Ca}^{2+}]_n$ was 685 ± 67 nM and $[\text{Ca}^{2+}]_c$ was 615 ± 73 nM for FPL 64176 treatment; $n = 31$). DNA transfection and RNase protection experiments demonstrate that FPL 64176-

induced Ca^{2+} transients stimulate expression of the endogenous *c-fos* gene and the transfected *c-fos* gene constructs pF711myc, pFos Δ CREmyc, pFos Δ SRFmyc, and pF222myc (Fig. 3a), although the response is weaker than that induced by KCl/FPL treatment (for example, in Fig. 3a, compare lanes 3 and 4). Ca^{2+} transients following KCl stimulation ($[\text{Ca}^{2+}]_n$ was 227 ± 12 nM; $[\text{Ca}^{2+}]_c$ was

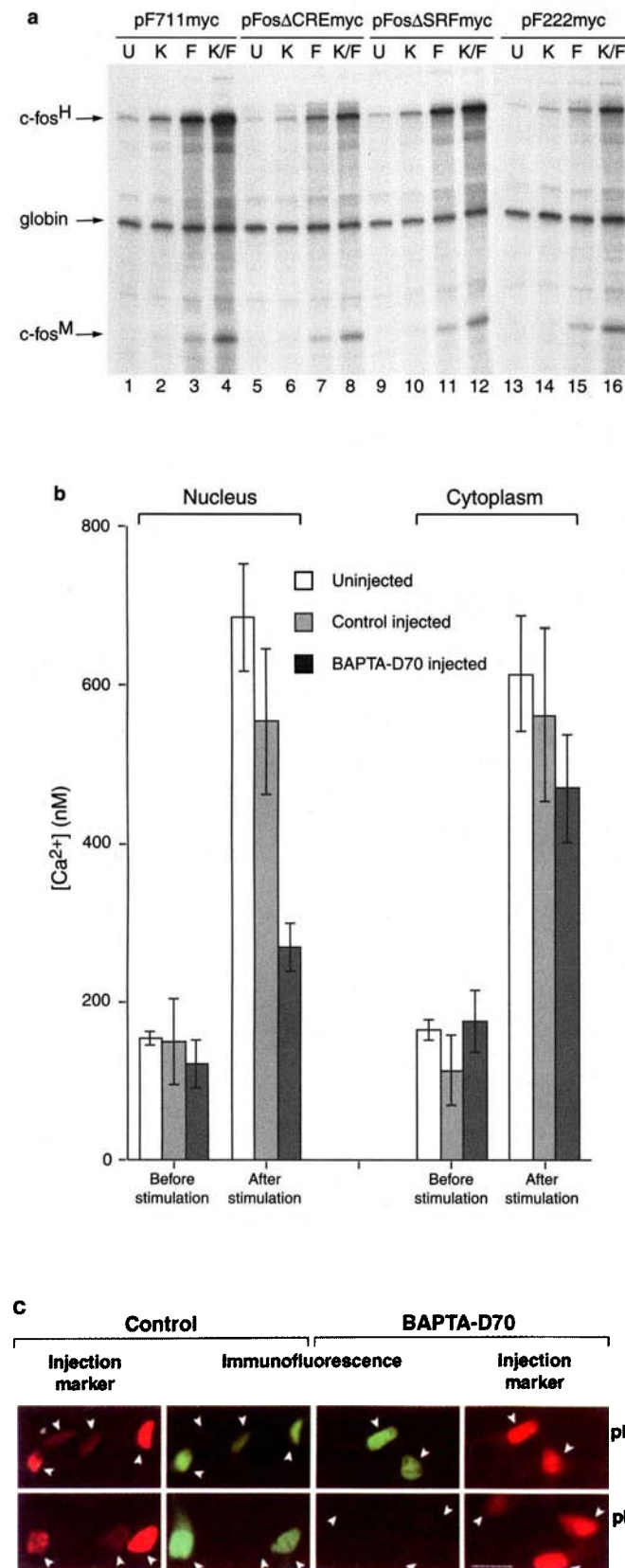


Figure 3 SRE-mediated Ca^{2+} -activated transcription can be triggered in the cytoplasm and does not require an increase in nuclear Ca^{2+} . **a**, RNase protection analysis of expression of the indicated transfected *c-fos* gene constructs in unstimulated cells (U) or cells stimulated with KCl (K), FPL 64176 (F), or KCl/FPL (K/F). Levels of *c-fos* mRNA transcribed from the indicated plasmids, normalized to the amount of α -globin mRNA, expressed as a percentage of the KCl/FPL-induced mRNA levels produced by the same plasmid are (mean $\pm$ s.e.m.): pF711myc, 19 ± 1 , $n = 5$ (K), 53 ± 9 , $n = 4$ (F); pFos Δ CREmyc, 20 ± 2 , $n = 5$ (K), 53 ± 6 , $n = 5$ (F); pFos Δ SRFmyc, 17 ± 3 , $n = 4$ (K), 52 ± 10 , $n = 4$ (F); pF222myc, 16 ± 2 , $n = 3$ (K), 37 ± 5 , $n = 3$ (F). **b**, Summary of Ca^{2+} imaging experiments. Nuclear microinjection of BAPTA-D70 ($n = 7$) reduced increases in $[\text{Ca}^{2+}]_n$ but not in $[\text{Ca}^{2+}]_c$ following FPL 64176 treatment, compared to uninjected cells ($n = 31$) ($P < 0.01$, two-tailed, unpaired *t*-test; $P < 0.01$, Mann-Whitney *U* test) and control-injected cells ($n = 6$) ($P < 0.01$, two-tailed, unpaired *t*-test; $P < 0.07$, Mann-Whitney *U* test). Resting $[\text{Ca}^{2+}]_n$ and $[\text{Ca}^{2+}]_c$ are not affected by nuclear microinjection of BAPTA-D70 (two-tailed, unpaired *t*-test and Mann-Whitney *U* test). **c**, Immunocytochemical analysis of FPL-induced expression of the indicated *c-fos* gene constructs microinjected into the nucleus together with BAPTA-D70 (right panels) or control solution (left panels). The Texas red fluorescence image identifies injected cells; arrowheads indicate cells microinjected into the nucleus. In cells microinjected into the nucleus with BAPTA-D70, the FPL-induced response is (mean $\pm$ s.e.m.) $83 \pm 8\%$ (pFos Δ CREmyc; 77 cells) and $19 \pm 2\%$ (pFos Δ SRFmyc; 61 cells) of that obtained in control-injected cells. Scale bar, $20 \mu\text{m}$. **d**, CRE-dependent (pFos Δ SRFmyc; circles) and SRE-dependent (pFos Δ CREmyc; squares) Ca^{2+} -activated transcription is expressed as a function of $[\text{Ca}^{2+}]_n$ in control-injected cells (open symbols) and cells injected into the nucleus with BAPTA-D70 (filled symbols). The absolute levels of CRE- and SRE-dependent expression in FPL-stimulated cells are $\sim 50\%$ of those in KCl/FPL-stimulated cells (see **a**).

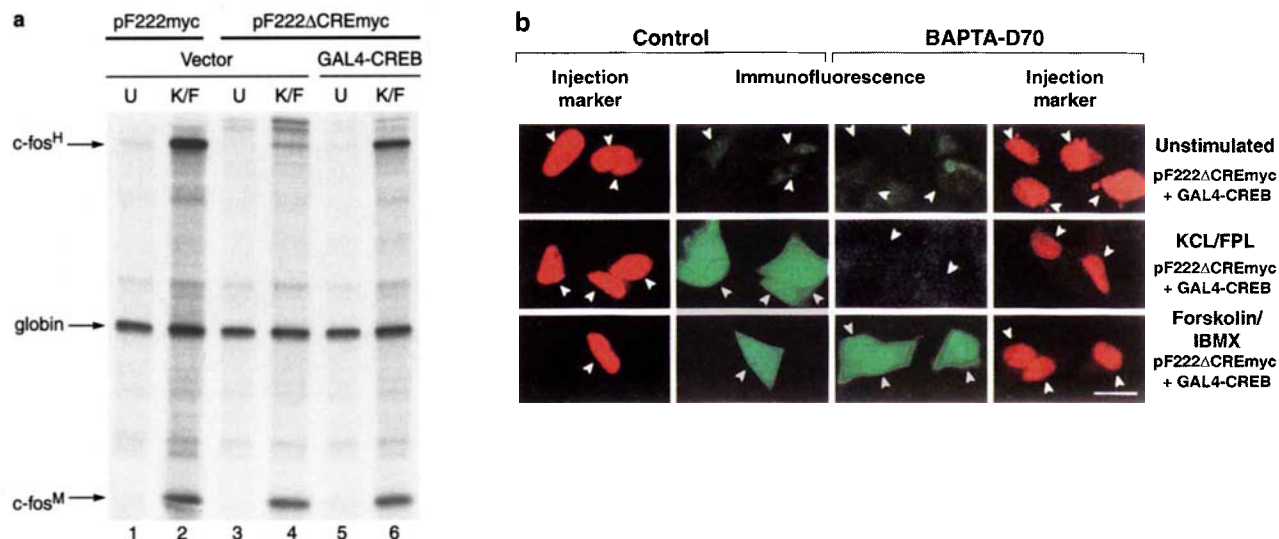


Figure 4 CREB functions as a nuclear calcium-responsive transcription factor. **a**, RNase protection analysis of expression of the indicated transfected *c-fos* gene constructs in unstimulated cells (U) or in cells stimulated with KCl/FPL (K/F). Cells were also transfected with a Gal4-CREB expression plasmid (lanes 5 and 6) or with the vector control (lanes 1–4). Expression of Gal4-CREB restores the KCl/FPL-induced transcriptional response of plasmid pF222ΔCREmyc to ~65–75% of that of plasmid pF222myc. **b**, Immunocytochemical analysis of Gal4-CREB-mediated expression of plasmid pF222ΔCREmyc microinjected into the nucleus

238 ± 12 nM; *n* = 48) did not efficiently activate transcription (Fig. 3a, lanes 2, 6, 10, 14).

To determine whether SRE-dependent transcriptional responses can be triggered by an increase in $[Ca^{2+}]_c$, we injected BAPTA-D70 into the nucleus which, upon FPL 64176 treatment, allowed $[Ca^{2+}]_c$ to rise to 470 ± 69 nM, but limited the increase in $[Ca^{2+}]_n$ to 269 ± 31 nM (*n* = 7) (Fig. 3b). This increase in $[Ca^{2+}]_n$ is similar to that which follows KCl treatment and as such is below the threshold for efficient activation of transcription (Fig. 3a). Nuclear microinjection of BAPTA-D70 did not affect FPL 64176-induced expression of pFosΔCREmyc (Fig. 3c, top panel; Fig. 3d), demonstrating that increases in $[Ca^{2+}]_c$ can control Ca^{2+} -dependent transcription mediated by the SRE. In contrast, FPL 64176-induced expression of pFosΔSRFmyc was blocked by nuclear microinjection of BAPTA-D70 (Fig. 3c, lower panel; Fig. 3d), confirming that an increase in $[Ca^{2+}]_n$ is critical for CRE-dependent Ca^{2+} -activated transcription. It remains to be seen whether an increase in $[Ca^{2+}]_n$ is also sufficient to stimulate CRE-dependent transcription.

We next investigated the role of the CRE-binding protein CREB in transcriptional regulation by nuclear Ca^{2+} signals. Expression of a Gal4-CREB fusion protein¹² confers KCl/FPL inducibility on plasmid pF222ΔCREmyc, which contains a Gal4-binding site in place of the CRE at nucleotide position -60 (Fig. 4a; compare lanes 4 and 6). Nuclear microinjection of BAPTA-D70 blocked the Gal4-CREB-mediated response following KCl/FPL treatment but had no effect on cAMP-induced transcriptional activation by Gal4-CREB (Fig. 4b). These experiments identify CREB as the first nuclear calcium-responsive transcription factor.

Taken together, we have shown that nuclear and cytoplasmic Ca^{2+} control transcription by distinct mechanisms. This finding provides a basis for differential gene regulation by neuronal activity that could, for example, explain differences in the transcriptional response when two different patterns of electrical stimulation are used to induce long-term potentiation in hippocampal neurons¹³. Cytoplasmic mechanisms, possibly involving Ras/MAP kinase signalling pathways^{14–16}, may be activated by brief electrical stimulations, whereas activation of nuclear enzymes involved in

alongside BAPTA-D70 (right panels) or control solution (left panels). The Texas red fluorescence image identifies injected cells; arrowheads indicate cells microinjected into the nucleus. Cells were stimulated with KCl/FPL, forskolin/IBMX, or were unstimulated. In cells injected into the nucleus with BAPTA-D70, the KCl/FPL-induced and forskolin/IBMX-induced response is (mean ± s.e.m.) 9.9 ± 0.3% (38 cells) and 104 ± 26% (64 cells), respectively, of that of control-injected cells. Scale bar, 20 μm.

transcriptional control, such as Ca^{2+} /calmodulin-dependent protein kinase IV^{17–20}, may require stronger or prolonged depolarization that gives rise to larger nuclear Ca^{2+} transients. In addition, Ca^{2+} entry through different types of Ca^{2+} channels^{11,21} may generate spatially distinct Ca^{2+} signals that determine exactly which genes are being activated. In conclusion, the intracellular localization of Ca^{2+} transients profoundly influences the transcriptional response and provides a previously unrecognized level of control that may determine the biological response to particular stimuli. □

Methods

Microinjection and DNA transfection. AtT20 cells were injected using a Zeiss Microinjection Workstation with either BAPTA-D70 injection solution (12.5% (w/v) BAPTA-D70 (Molecular Probes), half-strength PBS, pH 7.2, 4 mM $CaCl_2$, 1 mM $MgCl_2$, 0.9% (w/v) 70K dextran coupled to Texas red (Molecular Probes) or with the same solution without $CaCl_2$ and BAPTA-D70 substituted for neutral 70K dextran (control). The estimated BAPTA-D70 concentration in the nucleus is 0.8–2.4 mM. Plasmids were injected into the nucleus at a concentration of 0.5 mg ml⁻¹ in the injection solution. Cells were transfected using Lipofectamine (Gibco/BRL).

Calcium imaging. AtT20 cells were imaged 3–4 h after microinjection on poly-D-lysine-coated glass coverslips in a buffered salt-glucose-glycine (SGG) solution¹¹ at room temperature. Coverslips were mounted in a perfusion chamber and localization of the Texas red injection marker was recorded by taking an image of Texas red fluorescence (excitation 514 nm, emission ≥550 nm) on an MRC 600 confocal laser scanning microscope. Cells were loaded with 11 μM Fluo-3 AM (from a stock solution of 2.2 mM Fluo-3 dissolved in anhydrous DMSO containing 20% (w/v) Pluronic detergent) for 20 min, and then washed with SGG. Fluo-3 fluorescence images (excitation 488 nm, emission ≥515 nm) were taken before and after KCl, or FPL, or KCl/FPL stimulation. To calibrate the images, Fluo-3 was saturated by adding 50 μM ionomycin to the perfusion solution and then quenched with 10 mM $MnCl_2$ + 50 μM ionomycin to levels corresponding to ~100 nM Ca^{2+} (ref. 22). The K_d of Fluo-3 taken to be 315 nM and free Ca^{2+} concentrations were calculated according to the equation $[Ca^{2+}] = K_d(F - F_{min})/(F_{max} - F)$, where F is the fluorescence²³.

Plasmids. The human *c-fos* gene constructs used were derivatives of plasmids pF711 and pF222 (ref. 24). An oligonucleotide encoding the 9E10 Myc epitope (EQKLISEEDL, in single-letter code) was inserted in-frame into the fourth exon of the *c-fos* gene. The CRE was replaced by a Gal4-binding site. Mutation of the SRE²⁵, the Gal4-CREB expression vector¹², the BAG vector²⁶ and plasmid pSVα1²⁷ have been described.

Stimulations and gene expression analysis. Cells were depolarized by adding to the medium either 0.41 volumes of KCl depolarization solution (10 mM HEPES, pH 7.2, 170 mM KCl, 1 mM MgCl₂, 2 mM CaCl₂) (KCl stimulation), or 5 μM FPL 64176 (FPL stimulation), or 0.41 volumes of KCl depolarization solution containing 5 μM FPL 64176 (KCl/FPL stimulation). To increase intracellular levels of cAMP, cells were exposed to 10 μM forskolin (Sigma) and 0.5 mM IBMX (3-isobutyl-1-methylxanthine; Sigma). Expression of the endogenous *c-fos* protein and the Myc-tagged *c-fos* protein was analysed 2 h after stimulation using a rabbit polyclonal c-Fos-specific antibody (Santa Cruz) and the 9E10 monoclonal antibody, respectively, and standard biotin-avidin (fluorescein-conjugated) immunostaining. Expression of β-galactosidase was analysed similarly using a β-galactosidase-specific polyclonal antibody (5'-3', Boulder, Colorado). Immunofluorescence was quantified using an MRC 600 confocal laser scanning microscope. RNA levels were analysed 50 min after stimulation using an RNase protection assay^{11,27}.

Received 23 September; accepted 25 November 1996.

- Sheng, M. & Greenberg, M. E. *Neuron* **4**, 477-485 (1990).
- Morgan, J. I. & Curran, T. *Annu. Rev. Neurosci.* **14**, 421-451 (1991).
- Ghosh, A. & Greenberg, M. E. *Science* **268**, 239-247 (1995).
- Berridge, M. J. *Nature* **361**, 315-325 (1993).
- Clapham, D. E. *Cell* **80**, 259-268 (1995).
- Segal, M. & Mannor, D. J. *Physiol.* **448**, 655-676 (1992).
- Birch, B. D., Eng, D. L. & Kocsis, J. D. *Proc. Natl Acad. Sci. USA* **89**, 7978-7982 (1992).
- Al-Mohanna, F. A., Caddy, K. W. T. & Bolsover, S. R. *Nature* **367**, 745-750 (1994).
- O'Malley, D. M. *J. Neurosci.* **14**, 5741-5758 (1994).
- Zheng, W., Rampe, D. & Triggle, D. J. *Mol. Pharmacol.* **40**, 734-741 (1991).
- Bading, H., Ginty, D. D. & Greenberg, M. E. *Science* **260**, 181-186 (1993).
- Sheng, M., Thompson, M. A. & Greenberg, M. E. *Science* **252**, 1427-1430 (1991).
- Worley, P. F. et al. *J. Neurosci.* **13**, 4776-4786 (1994).
- Bading, H. & Greenberg, M. E. *Science* **253**, 912-914 (1991).
- Rosen, L. B., Ginty, D. D., Weber, M. J. & Greenberg, M. E. *Neuron* **12**, 1207-1221 (1994).
- Rusanescu, G., Qi, H., Thomas, S. M., Brugge, J. S. & Halegoua, S. *Neuron* **15**, 1415-1425 (1995).
- Jensen, K. F., Ohmsted, C. A., Fisher, R. S. & Sahyoun, N. *Proc. Natl Acad. Sci. USA* **88**, 2850-2853 (1991).
- Matthews, R. P. *Mol. Cell. Biol.* **14**, 6107-6116 (1994).
- Sun, P., Enslin, H., Myung, P. S. & Maurer, R. A. *Genes Dev.* **8**, 2527-2539 (1994).
- Miranti, C. K., Ginty, D. D., Huang, G., Chatila, T. & Greenberg, M. E. *Mol. Cell. Biol.* **15**, 3672-3684 (1995).
- Ghosh, A., Carnahan, J. & Greenberg, M. E. *Science* **263**, 1618-1623 (1994).
- Minta, A., Kao, J. P. Y. & Tsien, R. Y. *J. Biol. Chem.* **264**, 8171-8178 (1989).
- Tsien, R. Y., Pozzan, T. & Rink, T. J. *J. Cell Biol.* **94**, 325-334 (1982).
- Treisman, R. *Cell* **42**, 889-902 (1985).
- Hill, C. S. & Treisman, R. *EMBO J.* **14**, 5037-5047 (1995).
- Price, J., Turner, D. & Cepko, C. *Proc. Natl Acad. Sci. USA* **84**, 156-160 (1987).
- Sheng, M., Dougan, S. T., McFadden, G. & Greenberg, M. E. *Mol. Cell. Biol.* **8**, 2787-2796 (1988).

Acknowledgements: We thank M.J. Shipton for AtT20 cells, L. Lagnado for help with Ca²⁺ imaging, A. Randall for FPL 64176, and W. Wisden and R. Treisman for comments on the manuscript. This work was supported by the MRC and Glaxo/Wellcome.

Correspondence and requests for materials should be addressed to H.B. (e-mail: hbl@mrc-lmb.cam.ac.uk).

Actin polymerization is induced by Arp2/3 protein complex at the surface of *Listeria monocytogenes*

Matthew D. Welch*, Akihiro Iwamatsu† & Timothy J. Mitchison*

* Department of Cellular and Molecular Pharmacology, University of California at San Francisco, San Francisco, California 94143, USA

† Central Laboratories for Key Technology, Kirin Brewery Co., Yokohama, Japan

The pathogenic bacterium *Listeria monocytogenes* is capable of directed movement within the cytoplasm of infected host cells. Propulsion is thought to be driven by actin polymerization at the

bacterial cell surface^{1,2}, and moving bacteria leave in their wake a tail of actin filaments³. Determining the mechanism by which *L. monocytogenes* polymerizes actin may aid the understanding of how actin polymerization is controlled in the cell. Actin assembly by *L. monocytogenes* requires the bacterial surface protein ActA^{4,5} and protein components present in host cell cytoplasm. We have purified an eight-polypeptide complex that possesses the properties of the host-cell actin polymerization factor. The pure complex is sufficient to initiate ActA-dependent actin polymerization at the surface of *L. monocytogenes*, and is required to mediate actin tail formation and motility. Two subunits of this protein complex are actin-related proteins (ARPs) belonging to the Arp2 and Arp3 subfamilies. The Arp3 subunit localizes to the surface of stationary bacteria and the tails of motile bacteria in tissue culture cells infected with *L. monocytogenes*; this is consistent with a role for the complex in promoting actin assembly *in vivo*. The activity and subunit composition of the Arp2/3 complex suggests that it forms a template that nucleates actin polymerization.

The intracellular motility of *L. monocytogenes* resembles the protrusion of the lamellipodia of locomoting cells in that propulsion is thought to be driven by the polymerization of actin filaments². Actin assembly at the surface of *L. monocytogenes* is initiated by ActA, a bacterial surface protein that recruits host-cell actin polymerization factors. We purified these host-cell factors from human platelets, which are highly enriched in actin polymerization proteins. To further enrich for these factors, we prepared a soluble extract from the triton-insoluble cytoskeleton of platelets by treatment with 0.6 M KCl. This 'cytoskeleton extract' supported actin tail formation by *L. monocytogenes* (Fig. 1A) and motility at rates (mean = $3.8 \pm 0.2 \mu\text{m min}^{-1}$) that were approximately half of those observed in platelet 'cytoplasmic extract' (mean = $6.5 \pm 0.3 \mu\text{m min}^{-1}$) and extracts from *Xenopus laevis* eggs (mean = $7.0 \pm 0.3 \mu\text{m min}^{-1}$)⁶.

Actin polymerization-promoting factors were purified from cytoskeleton extract by column chromatography. Fractions were tested for their ability to promote *L. monocytogenes*-mediated actin assembly by adding DAPI-labelled bacteria and rhodamine-labelled actin. Active fractions were capable of promoting the assembly of actin 'clouds' and/or actin 'tails' associated with bacteria, as visualized by fluorescence microscopy. Incubation of *L. monocytogenes* with actin alone did not result in actin association with bacteria.

Three sequential chromatographic steps were used to purify the actin assembly-promoting activity (Table 1). Cytoskeleton extract (Fig. 2a, lane 1) was passed over a Q-sepharose column and the activity collected in the flow-through fraction (Fig. 2a, lane 2). The Q-sepharose flow-through was then chromatographed on an S-sepharose column. Fractions from S-sepharose that corresponded to the peak of activity contained eight major polypeptides (Fig. 2a, lane 3). To determine which of the proteins were responsible for the observed activity, the S-sepharose peak was further fractionated by Superose-6 gel filtration chromatography (Fig. 2a lane 4, 2b). The elution profile of the activity from the gel filtration column corresponded exactly to the elution profile of a complex of eight polypeptides (Fig. 2c); minor contaminating proteins did not co-fractionate with the activity. All eight polypeptides also co-fractionated by sucrose-gradient sedimentation (data not shown), indicating that they form a stable complex.

The proteins in this complex are present in roughly equal stoichiometry, based on quantification of Coomassie-dye binding (except for the substoichiometric subunits of relative molecular mass (M_r) 39K and 40K, which when quantified as a unit are stoichiometric with the others). Summation of the molecular masses of the subunits determined by SDS-PAGE (polyacrylamide gel electrophoresis) predicted a total M_r of 231K. Using values for its Stoke's radius (5.5 nm) and sedimentation coefficient (9.5 S), the native M_r of the complex was estimated to be 216K (ref. 7), indicating that it consists of one of each polypeptide.

The purified complex is sufficient to promote the assembly of dense actin 'clouds' associated with bacteria (Fig. 1Ba–c). These clouds generally surrounded (Fig. 1Ba) or emanated from one end of a bacterium (Fig. 1Bb). Control bacteria with an in-frame deletion of the *actA* gene did not exhibit any actin assembly at their surface in the presence of the purified complex (Fig. 1Bd). This is consistent with the requirement for ActA in *L. monocytogenes*-mediated actin assembly *in vivo*^{4,5}.

In rare instances (<1% of bacteria observed), the pure complex could promote the assembly of very short actin tails (Fig. 1Bc). However, motility assays showed little or no movement. Addition of the Q-binding fraction (from the first step in the purification procedure) to the pure complex resulted in a five- to tenfold stimulation of cloud formation, a similar enhancement of actin tail formation (Fig. 1C; to 5% of bacteria observed), and a partial restoration of motility (to 0.1–0.5 $\mu\text{m min}^{-1}$). The Q-binding fraction alone had no actin assembly-promoting activity (data not shown), indicating that the complex of eight polypeptides is required to mediate actin tail formation and motility.

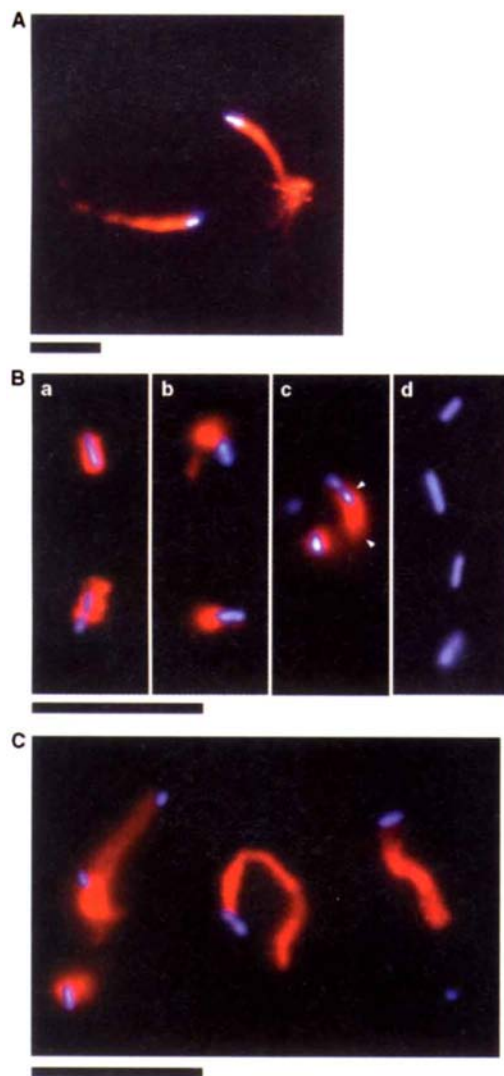


Figure 1 Actin structures formed by *L. monocytogenes* in: **A**, the cytoskeleton extract (the starting material for purification); **B**, the pure eight-polypeptide complex (**a**, **b**, actin clouds assembled by bacteria; **c**, short actin tails (delimited by the arrowheads emanating from a bacterium); **d**, bacteria carrying a deletion of the *actA* gene); **C**, the pure complex plus the Q-binding fraction (equal volumes of each). Bacteria are labelled with DAPI (blue), and actin with rhodamine (red). Scale bars, 10 μm .

We determined the molecular identity of selected subunits of the purified protein complex by peptide sequencing. The 43K and 50K subunits were identified as actin-related proteins (ARPs)⁸ belonging to the Arp2 and Arp3 subfamilies, respectively (Fig. 3a). The 43K protein is most similar to chicken Arp2 (ACTL)⁹, and the 50K protein is most similar to bovine Arp3 (actin2)¹⁰. We have therefore named the purified complex the Arp2/3 complex. The native molecular mass (~220K) and polypeptide composition of the human Arp2/3 complex are strikingly similar to a seven-protein complex purified from *Acanthamoeba castellanii* by profilin affinity chromatography¹¹ (Fig. 3b). These similarities suggest that the structure and function of the Arp2/3 complex has been conserved through evolution.

The biochemical activity of the Arp2/3 complex suggests that it should localize to regions where actin polymerization is induced by *L. monocytogenes* in infected cells. To test this assertion we generated antibodies to a peptide of 19 amino acids that corresponds to the C

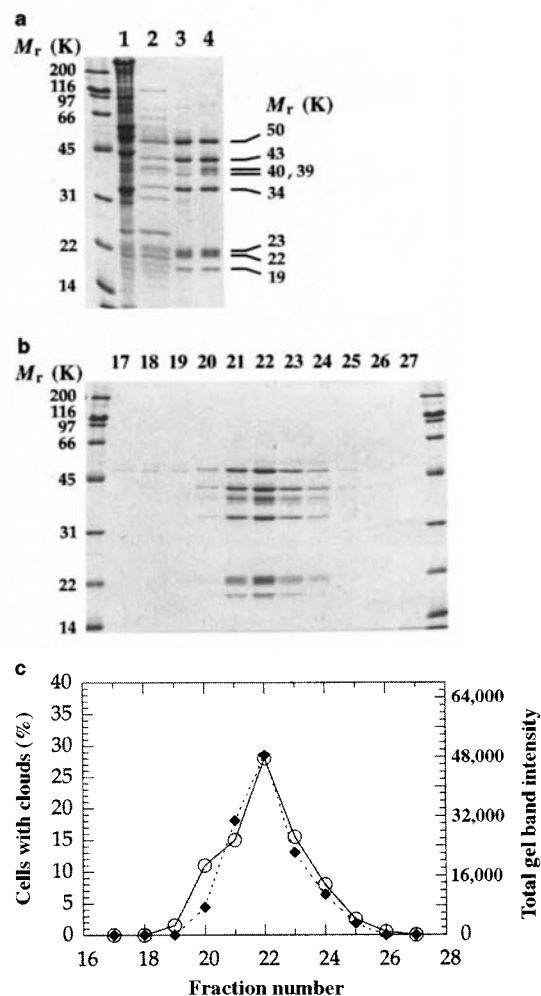


Figure 2 Purification of the factor from human platelets that promotes actin assembly at the surface of *L. monocytogenes*. **a**, The protein composition of fractions from the purification are visualized by 13% SDS-PAGE and staining with Coomassie blue. Lane 1, cytoskeleton extract (30 μg); lane 2, Q-sepharose flow-through (6 μg); lane 3, S-sepharose peak (6 μg); lane 4, Superose-6 peak (6 μg). **b**, The protein composition of the peak fractions (17–27) from the Superose-6 column are visualized by 13% SDS-PAGE and staining with Coomassie blue. **c**, The activity in superose-6 fractions 17–27 (measured by counting the percentage of bacteria that assembled a visible actin cloud or tail) is shown by the open circles. The relative amounts of the eight-protein complex in each fraction (determined by densitometry) is shown by the filled diamonds.

Table 1 Purification table

Fraction	Activity (units)	Protein (mg)	Specific activity (units mg ⁻¹)	Yield (%)
Cytoskeleton extract	8.0 × 10 ⁴	91	880	100
Q flow-through	1.6 × 10 ⁴	7.3	2,200	20
S peak	1.8 × 10 ⁴	1.7	10,600	23
Superose-6 peak	2.0 × 10 ³	0.6	3,300	3

One unit of activity per microlitre was defined as the concentration at which actin structures were observed to be associated with 15% of bacteria in the actin-assembly assay. A fraction exhibiting more than one unit of activity was diluted to a concentration (1/*x* its original concentration) at which 15% of bacteria assembled actin structures. This fraction was assigned *x* units of activity. The loss of both specific activity and total activity during the gel-filtration step may be due to the inherent instability of the activity (*t*_{1/2} < 1 day at 4°C), or to separation of the activity from a stimulatory factor present at low stoichiometry relative to the purified complex.

terminus of the Arp3 subunit. Affinity-purified antibodies recognized Arp3 in the purified complex (data not shown) and a single 50K protein in HeLa cell extracts (Fig. 3c). By immunofluorescence, we found that Arp3 localized to the surface of stationary bacteria and throughout the tails of moving bacteria (Fig. 3e), similar to the distribution of actin (Fig. 3d). However, Arp3 was not found in the stress fibres of HeLa cells (data not shown), indicating that it is only present in a subset of actin structures. Because Arp3 is tightly

associated in a protein complex, the distribution of Arp3 is likely to mirror that of the entire Arp2/3 complex.

The activity and localization of the Arp2/3 complex suggest that it represents the factor that initiates actin polymerization at the surface of *L. monocytogenes* *in vivo*. A model depicting how the Arp2/3 complex might promote actin polymerization at the bacterial surface is shown in Fig. 4. Because the activity of the Arp2/3 complex requires ActA expression, the complex must be recruited to the bacterial surface by interacting at least transiently with ActA. Binding to ActA could activate the complex, restricting actin polymerization to the bacterial surface, as has been observed *in vivo*^{1,2}. Active complex may promote actin polymerization by serving as a template for the nucleation of actin assembly^{11,12} (Fig. 4). Atomic modelling of Arp2 and Arp3 proteins¹² suggested that the complex may reside at the pointed end of an actin filament

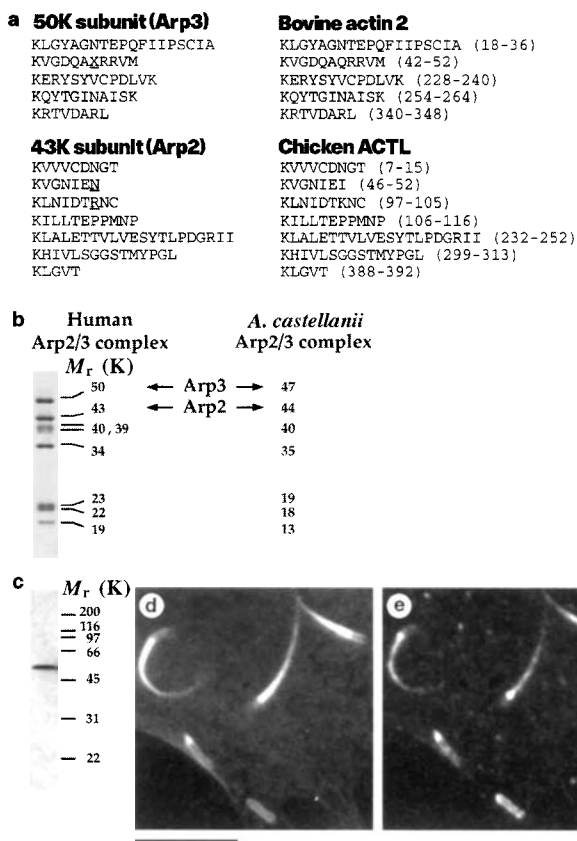


Figure 3 Identification and intracellular localization of subunits in the actin assembly-promoting complex. **a**, Peptide sequences from the human 50K and 43K proteins (left) are similar to the predicted amino-acid sequences from bovine actin2 (Arp3)¹⁰ and chicken ACTL (Arp2)⁹ (right). Where there are differences, residues in the peptide sequence are underlined. An X in the peptide sequence indicates that the amino acid could not be determined. The residue numbers of the bovine and chicken sequences are listed in parentheses. **b**, A comparison of the subunit composition of the human (this study) and *A. castellanii*¹¹ Arp2/3 complexes. **c**, Immunoblot of HeLa whole-cell extract probed with anti-Arp3 antibody. **d**, Distribution of actin in a HeLa cell infected with *L. monocytogenes*, visualized by indirect immunofluorescence. **e**, Distribution of Arp3 in the same cell, visualized by indirect immunofluorescence. The observed localization was competed by the inclusion of the Arp3 C-terminal peptide in the primary antibody solution, indicating that the staining is specific. Scale bar, 10 µm.

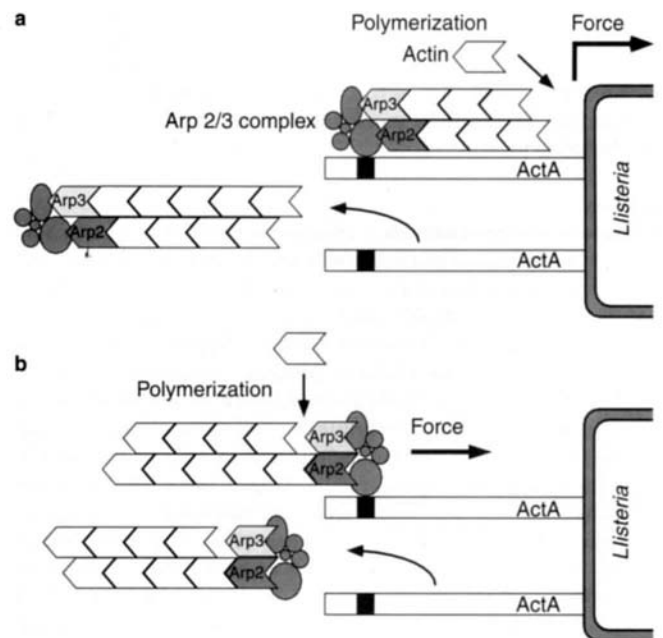


Figure 4 Models for Arp2/3 complex nucleation of actin-filament polymerization at the surface of *L. monocytogenes*. The Arp2/3 complex promotes actin polymerization by interacting with ActA at a site (amino acids 128-152) that is essential for actin assembly³¹ and bacterial motility²⁰. Actin polymerization initiated by the Arp2/3 complex contributes to the force that drives bacterial propulsion. After nucleating filament assembly the Arp2/3 complex and its associated actin filament dissociate from ActA and become incorporated into the actin tail. Actin is shown as a chevron to indicate the barbed and pointed ends of the actin monomer and filament. The models differ in the nature of the association between the Arp2/3 complex and the actin filament. **a**, The Arp2/3 complex binds to the pointed end of an actin filament and nucleates barbed-end polymerization¹². **b**, The Arp2/3 complex binds to the barbed end of an actin filament and remains associated with the filament end while nucleating barbed-end polymerization.

and nucleate barbed-end polymerization (Fig. 4a) by a mechanism similar to the nucleation of microtubule assembly by the γ -tubulin ring complex^{13,14}. Alternatively, the complex may both reside at and promote polymerization onto the barbed end of a filament, an activity first postulated for insertin^{15,16} (Fig. 4b). After initiating actin assembly, the Arp2/3 complex and its associated filament may dissociate from ActA and become incorporated into the actin tail, in agreement with our immunofluorescence data.

The activity of the Arp2/3 complex at the *L. monocytogenes* surface suggests that its cellular function may be to promote actin polymerization. Consistent with this assertion, the subunits of the *A. castellanii* Arp2/3 complex are localized to the filopodia and cell cortex of amoebae, which are regions of dynamic actin-filament assembly^{11,12}. Furthermore, the *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* Arp2 and Arp3 proteins are localized to cortical actin structures, and mutations in their corresponding genes cause defects in actin function^{17,18}. The relative molar amounts of Arp3:actin (1:60 in platelets (M.D.W. and T.J.M., unpublished results), 1:100 in *A. castellanii*¹²) and Arp2:actin (1:40 in *A. castellanii*¹²) suggest that the Arp2/3 complex may represent an important actin-polymerizing activity in cells.

The Arp2/3 complex initiates actin polymerization at the *L. monocytogenes* surface and is required to mediate actin tail formation and motility. However, additional factors are needed to stimulate bacterial motility. These factors may include vasodilator-stimulated phosphoprotein and profilin, which are localized to the surface of moving *L. monocytogenes* and have been implicated in fast bacterial movement^{6,19–21}. Full reconstitution of *L. monocytogenes* motility with purified components will require the recapitulation of all aspects of actin polymerization dynamics by combining nucleation (Arp2/3 complex), filament elongation, filament cross-linking and filament depolymerizing activities. This may represent the most promising approach towards elucidating the biochemical mechanisms that underlie more complex cellular processes, such as lamellipodial protrusion. □

Methods

Preparation of platelet extracts. To prepare the cytoskeleton extract, freshly outdated human platelet enriched plasma was obtained from a blood bank. Platelets were isolated by centrifugation²², washed three times in wash buffer (in mM: 20 PIPES, pH 6.8, 40 KCl, 5 EGTA, 1 EDTA) and resuspended in 1 vol wash buffer. Five vols of cold lysis buffer (wash buffer plus 0.05 mM ATP, 1% Triton X-100 and protease inhibitors (10 μ g ml⁻¹ leupeptin, pepstatin-A, chymostatin, 1 mM PMSF, 10 μ M benzamidin-HCl)) was added and the Triton-insoluble cytoskeleton was pelleted by centrifugation²². The cytoskeleton was resuspended in 5 vols of lysis buffer plus 100 mM sucrose, 0.2 mM DTT, and quick frozen for storage at -80°C . Thawed Triton cytoskeleton was extracted with 0.6 M KCl, depleted of actin²², and then desalted by gel filtration to produce the cytoskeleton extract. To prepare the cytoplasmic extract, platelets were isolated and washed as above, except that the composition of wash buffer was different (in mM: 20 MES, pH 6.1, 50 KCl, 2 MgCl₂, 10 EGTA, 1 EDTA, 0.5 ATP and 150 sucrose). Platelets were resuspended in 1 vol. lysis buffer (wash buffer plus protease inhibitors). Cells were lysed by sonicating. DTT was added to 0.2 mM and a clarified extract prepared by centrifugation at 100,000g for 1 h at 4°C . Extract was quick frozen for storage at -80°C .

Assay for activities that promote actin assembly at the surface of *L. monocytogenes*. Cytoskeleton extract or fractionated extract (10 μ l) was supplemented with *N*-hydroxysuccinimidyl 5-carboxytetramethyl rhodamine (TMR)-labelled actin²³ to 1 μ M, platelet actin²² to 1 μ M, ATP regenerating mix²⁴, Triton X-100 to 0.1%, catalase and glucose oxidase to 0.1 mg ml⁻¹, glucose to 10 mM, and DAPI-labelled *L. monocytogenes*. The assay mix was squashed between a glass microscope slide and coverslip, and incubated at 22°C for 20 min before viewing. DAPI-labelled *L. monocytogenes* were prepared from strain SLCC-5764 (ref. 25) and control strain DP-L1955, a derivative of SLCC-5764 with an in-frame deletion of the *actA* gene²⁶ (G. Smith and D. Portnoy, unpublished data). Saturated bacterial cultures grown in brain-heart infusion media at 37°C were supplemented with 10 μ g ml⁻¹ DAPI and further

incubated for 2.5 h. Bacteria were killed in culture with iodoacetic acid⁶. Cells were pelleted in a microfuge for 5 s, washed with buffer (in mM: 10 HEPES, pH 7.7, 50 KCl, 100 sucrose, 5 EGTA), resuspended in buffer plus 20% glycerol, and frozen in small portions at -80°C .

Purification of the Arp2/3 complex. Cytoskeleton extract in Q-buffer (in mM: 20 Tris, pH 8.0, 100 KCl, 2 MgCl₂, 5 EGTA, 1 EDTA, 0.5 DTT, 0.5 ATP, 2.5% glycerol) was passed over a 5-ml Hi-trap Q-sepharose HP column (Pharmacia) equilibrated with Q-buffer. The flow-through was supplemented with 40 mM MES, pH 6.1, 10% glycerol, diluted 1:1 into S-buffer (in mM: 20 MES, pH 6.1, 2 MgCl₂, 5 EGTA, 1 EDTA, 0.5 DTT, 0.5 ATP, 10% glycerol) and passed over a 1-ml Hi-trap SP-sepharose HP column (Pharmacia) equilibrated with S-buffer plus 50 mM KCl. The activity was eluted at 200 mM KCl with a linear salt gradient of 50–500 mM KCl. Peak fractions were pooled, concentrated using a centricon-30 (Amicon), and run over a Superose-6 HR 10/30 column (Pharmacia) equilibrated with (in mM) 20 MOPS, pH 7.0, 2 MgCl₂, 5 EGTA, 1 EDTA, 0.5 DTT, 0.5 ATP, 2.5% glycerol.

Determination of the physical properties of the complex. The Stoke's radius of the protein complex was estimated by Superose-6 gel filtration chromatography. The sedimentation coefficient was determined by 5–20% sucrose gradient sedimentation.

Peptide sequencing. Proteins were blotted to PVDF membrane (Applied Biosystems) and identified by staining with Ponceau S²⁷. The PVDF-immobilized proteins were reduced, S-carboxymethylated, and digested *in situ* with Achromobacter protease I and endoproteinase Asp-N²⁸. Digested peptides were chromatographed by reverse-phase high-performance liquid chromatography (HPLC) using a Wakosil-II AR C18 300 Å column (Wako Pure Chemical). Amino-acid sequencing was performed with a gas-phase sequencer (model PPSQ-10, Shimadzu).

Immunoblotting and immunofluorescence. Polyclonal anti-peptide antibodies that recognize human Arp3 were generated by immunizing rabbits with the synthetic peptide CYEEIGPSIVRHNPVFGVMS, corresponding to the last 19 amino acids of bovine Arp3. An N-terminal cysteine was added to the peptides for sulphhydryl coupling. Peptide conjugation and antibody affinity purification were performed²⁹. Antibodies were concentrated by dialysis into 10 mM MOPS, pH 6.8, 100 mM NaCl, 35% glycerol. Immunoblotting was performed with affinity-purified anti-Arp3 primary antibody (0.4 μ g ml⁻¹) using the ECL system (Amersham). For immunofluorescence experiments, HeLa cells were infected with *L. monocytogenes* strain 10403S⁶. Cells were extracted for 30 s in cytoskeleton buffer³⁰ with 1 μ g ml⁻¹ phalloidin, then fixed in cytoskeleton buffer plus 4% formaldehyde for 10 min followed by -20°C methanol for 3 min (methanol fixation was necessary to observe Arp3 staining, indicating that the antibody recognizes only denatured protein). Indirect immunofluorescence was performed using affinity-purified rabbit anti-Arp3 (4 μ g ml⁻¹) and mouse anti-actin (10 μ g ml⁻¹; Boehringer Mannheim) primary antibodies.

Received 1 August; accepted 21 November 1996.

- Sanger, J. M., Sanger, J. W. & Southwick, F. S. *Infect. Immun.* **60**, 3609–3619 (1992).
- Theriot, J. A., Mitchison, T. J., Tilney, L. G. & Portnoy, D. A. *Nature* **357**, 257–260 (1992).
- Tilney, L. G. & Portnoy, D. A. *J. Cell Biol.* **109**, 1597–1608 (1989).
- Kocks, C. *et al.* *Cell* **68**, 521–531 (1992).
- Domann, E. *et al.* *EMBO J.* **11**, 1981–1990 (1992).
- Theriot, J. A., Rosenblatt, J., Portnoy, D. A., Goldschmidt-Clermont, P. J. & Mitchison, T. J. *Cell* **76**, 505–517 (1994).
- Siegel, L. M. & Monty, K. J. *Biochim. Biophys. Acta* **112**, 346–362 (1966).
- Frankel, S. & Mooseker, M. *Curr. Opin. Cell Biol.* **8**, 30–37 (1996).
- Michaille, J. J., Gouy, M., Blanchet, S. & Duret, L. *Gene* **154**, 205–209 (1995).
- Tanaka, T. *et al.* *Biochem. Biophys. Res. Commun.* **187**, 1022–1028 (1992).
- Machesky, L. M., Atkinson, S. J., Ampe, C., Vandekerckhove, J. & Pollard, T. D. *J. Cell Biol.* **127**, 107–115 (1994).
- Kelleher, J. F., Atkinson, S. J. & Pollard, T. D. *J. Cell Biol.* **131**, 385–397 (1995).
- Zheng, Y., Wong, M. L., Alberts, B. & Mitchison, T. *Nature* **378**, 578–583 (1995).
- Moritz, M., Braunfeld, M. B., Sedat, J. W., Alberts, B. M. & Agard, D. A. *Nature* **378**, 638–640 (1995).
- Ruhnau, K., Gaertner, A. & Wegner, A. *J. Mol. Biol.* **210**, 141–148 (1989).
- Gaertner, A. & Wegner, A. *J. Muscle Res. Cell Motil.* **12**, 27–36 (1991).
- Moreau, V., Madania, A., Martin, R. P. & Winsor, B. J. *J. Cell Biol.* **134**, 117–132 (1996).
- McCollum, D., Feoktistova, A., Morpheu, M., Balasubramanian, M. & Gould, K. L. *EMBO J.* **15**, 6438–6446 (1996).
- Chakraborty, T. *et al.* *EMBO J.* **14**, 1314–1321 (1995).
- Lasa, I., Violaine, D., Gouin, E., Marchand, J. B. & Cossart, P. *Mol. Microbiol.* **18**, 425–436 (1995).
- Smith, G. A., Theriot, J. A. & Portnoy, D. A. *J. Cell Biol.* **135**, 647–660 (1996).
- Schaefer, S. R. *Methods Enzymol.* **215**, 58–77 (1992).
- Kellogg, D. R., Mitchison, T. J. & Alberts, B. M. *Development* **103**, 675–686 (1988).
- Murray, A. W. *Methods Cell Biol.* **36**, 581–605 (1991).
- Leimeister-Wachter, M. & Chakraborty, T. *Infect. Immun.* **57**, 2350–2357 (1989).

26. Brundage, R. A., Smith, G. A., Camilli, A., Theriot, J. A. & Portnoy, D. A. *Proc. Natl Acad. Sci. USA* **90**, 11890–11894 (1993).
27. Hirano, H. & Watanabe, T. *Electrophoresis* **11**, 573–580 (1990).
28. Iwamatsu, A. *Electrophoresis* **13**, 142–147 (1992).
29. Sawin, K. E., Mitchison, T. J. & Wordeman, L. G. *J. Cell. Sci.* **101**, 303–313 (1992).
30. Symons, M. H. & Mitchison, T. J. *J. Cell Biol.* **114**, 503–513 (1991).
31. Pistor, S., Chakraborty, T., Walter, U. & Wehland, J. *Curr. Biol.* **5**, 517–525 (1995).

Acknowledgements: We thank J. Barry, D. Morgan, K. Oegema and Y. Zheng for advice on protein purification; D. Portnoy and J. Theriot for bacterial strains; and J. Theriot and D. Portnoy and all members of the Mitchison lab for discussions and comments on the manuscript. M.D.W. is an NIH postdoctoral fellow. This work was supported by a grant from the NIH (T.J.M.).

Correspondence and requests for materials should be addressed to M.D.W. (e-mail: welch@cgl.ucsf.edu).

Extension of chromatin accessibility by nuclear matrix attachment regions

Thomas Jenuwein^{*†}, William C. Forrester^{*}, Luis A. Fernández-Herrero^{*}, Götz Laible[†], Maude Dull^{*} & Rudolf Grosschedl^{*}

^{*} Howard Hughes Medical Institute and Department of Microbiology and Immunology, University of California, San Francisco, California 94143-0414, USA

[†] Institute of Molecular Pathology, Dr Bohrgasse 7, A-1030 Vienna, Austria

Transcription of the variable region of the rearranged immunoglobulin μ gene is dependent on an enhancer sequence situated within one of the introns of the gene. Experiments with transgenic mice have shown that activation of the promoter controlling this transcription also requires the matrix-attachment regions (MARs) that flank the intronic enhancer¹. As this μ gene enhancer can establish local areas of accessible chromatin², we investigated whether the MARs can extend accessibility to more distal positions. We eliminated interactions between enhancer- and promoter-bound factors by linking μ enhancer/MAR fragments to the binding sites for bacteriophage RNA polymerases that were either close to or one kilobase distal to the enhancer. The μ enhancer alone mediated chromatin accessibility at the proximal site but required a flanking MAR to confer accessibility upon the distal promoter. This long-range accessibility correlates with extended demethylation of the gene construct but not with whether it is being actively transcribed. MARs thus collaborate with the μ enhancer to generate an extended domain of accessible chromatin.

Transcriptional enhancers and locus control regions stimulate transcription from promoters over long distances³. Although transcriptional stimulation requires interactions between enhancer- and promoter-binding proteins by DNA looping⁴, enhancer-mediated changes in chromatin structure have been proposed as an additional mechanism in promoter activation⁵. The murine immunoglobulin μ -gene enhancer region functions as a locus control region that directs lymphoid-specific high-level expression of a rearranged μ gene or of a heterologous gene in transgenic mice independently of its chromosomal position^{6,7}. The μ enhancer, consisting of a 95-base-pair (bp) core element and a promoter region that generates non-coding μ transcripts ($I\mu$), is flanked on either side by (A + T)-rich MARs^{8–10}. MARs, also referred to as SARs, are operationally defined as DNA sequences that associate with the nuclear matrix or scaffold^{11,12}. Functionally, MARs can augment the activity of flanking enhancers^{1,13} and protect genes from position effects by acting as boundary elements¹⁴.

In principle, MARs could augment μ enhancer function by propagating an enhancer-induced alteration in chromatin structure or by facilitating long-range interactions between enhancer- and promoter-bound factors. To investigate a possible role for MARs in

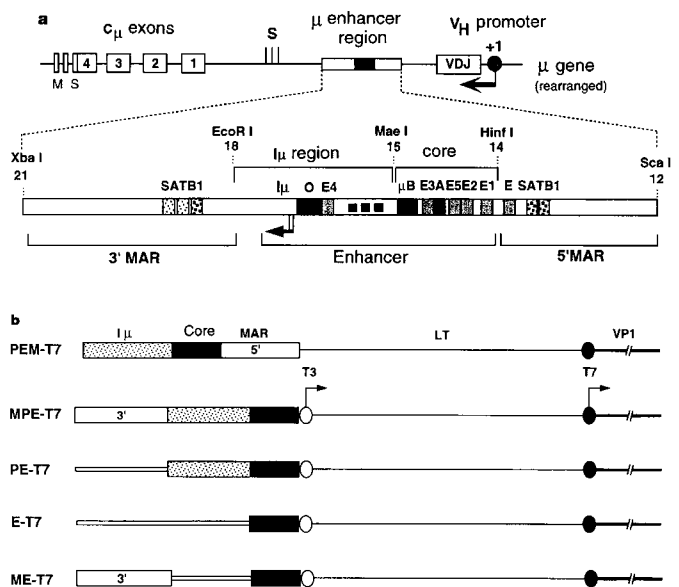


Figure 1 Structure of transgenes. **a**, The intronic murine immunoglobulin heavy-chain μ enhancer region, including the enhancer core (E), $I\mu$ promoter region (P) and flanking MARs (M). The numbers above refer to $\sim 1/100$ of the nucleotide positions relative to the transcription start site in the rearranged μ gene. Individual factor binding sites are indicated¹⁰. The μ enhancer region in the reporter gene constructs is inserted in the antisense orientation. **b**, The transgene constructs. The T3 and T7 promoters are indicated by circles (T3) and filled circles (T7). The PEM-T7 construct lacks the enhancer-proximal T3 promoter. The 1-kb LT and 1-kb VP1 reporter sequences, from the SV40 genome, are represented by thin and bold lines, respectively. 'Stuffer' sequences replacing individual enhancer components are indicated by an open bar. The direction of T3- and T7-specific transcription (thin arrows) is opposite to that of heterogeneously initiated transcription from the $I\mu$ promoter (thick arrow in **a**).

regulating chromatin accessibility, we introduced gene constructs into the mouse germ line, in which the μ enhancer, alone or together with a MAR, was linked to binding sites for bacteriophage RNA polymerases (Fig. 1). The bacteriophage T3 and T7 RNA polymerases recognize their promoters in the absence of protein-protein interactions and, therefore, T3- and T7-specific transcription in chromatin reflects access to polymerase¹⁵. In the gene constructs, we inserted a T3 promoter immediately adjacent to the μ enhancer, and a T7 promoter 1 kilobase (kb) downstream of the enhancer, mimicking the natural positions of the $I\mu$ promoter and the variable-region (VH) promoter in the rearranged μ gene. For each gene construct, we generated multiple transgenic mice and obtained pre-B cell lines by transformation with Abelson murine leukaemia virus².

To analyse factor access at the T3 and T7 promoters in native chromatin, we isolated nuclei from transgenic pre-B cells and incubated them with the respective RNA polymerases. We detected T3- and T7-specific transcripts by S1 nuclease protection assay with SV40 large-T(LT)-specific and VP1-specific probes, respectively (Fig. 2, middle and bottom panels). Transcripts initiating at the enhancer-proximal T3 promoter were found in all lines, consistent with previous work showing that the μ enhancer core was necessary and sufficient to establish local accessibility in chromatin². Moreover, the relative number of T3-specific transcripts were proportional to both the numbers of transgene copies and to T7-specific transcripts generated from naked genomic DNA (Fig. 2, top panel). We also detected abundant T7-specific transcripts in all five lines containing transgenes that include the μ enhancer core, the $I\mu$ promoter region, and a single MAR at either the 5' or 3' position. By contrast, access to T7 RNA polymerase in transgenes that include

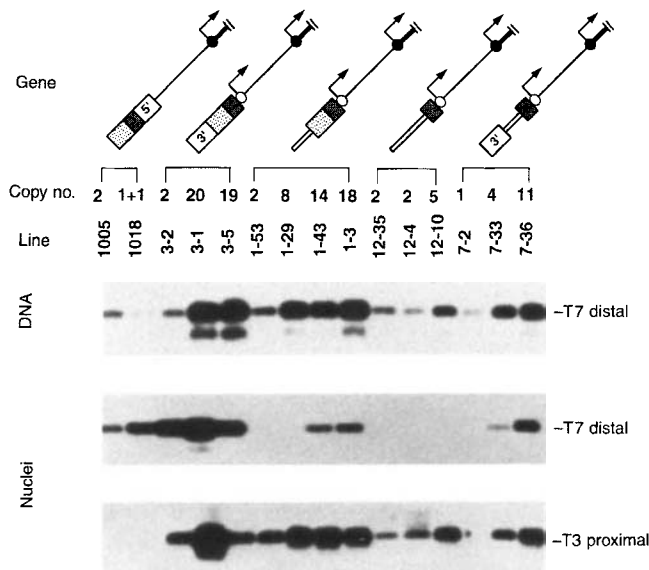


Figure 2 Analysis of enhancer-mediated short- and long-range factor access in nuclear chromatin. Top, S1 nuclease protection analysis of correctly initiated T7-specific P1 transcripts, generated by *in vitro* transcription of genomic pre-B cell DNA with T7RNA polymerase. Individual transgenic pre-B cell lines are arranged according to the type of gene construct and the copy number, which was determined by DNA blot analysis (not shown) and by *in vitro* transcription of genomic DNA with T7 RNA polymerase (this figure). Middle, S1 nuclease protection analysis of T7-specific VP1 transcripts, generated by incubation of transgenic pre-B cell nuclei with T7 RNA polymerase. Bottom, S1 nuclease protection analysis of T3-specific LT transcripts in transgenic pre-B cell nuclei. The transgenes present in lines 1005 and 1018 lack the enhancer-proximal T3 promoter and therefore were not used for the analysis of T3 transcription.

the enhancer core and $I\mu$ promoter, but lack both MARs, was reduced five- to ten-fold in two out of the four lines examined, and no T7-specific transcripts were detected in the other two lines. Transgenes carrying the μ enhancer core, but lacking the MARs and $I\mu$ promoter, failed to generate T7-specific transcripts in all three lines examined, whereas T3-specific transcripts were detected at moderate reduced levels. In combination with a single MAR, the μ enhancer core lacking the $I\mu$ promoter region mediated T7 accessibility in two out of three lines, albeit at five- to ten-fold lower levels relative to transgenes also containing the $I\mu$ promoter region. We confirmed the contribution of MARs to long-range chromatin accessibility by digestion of nuclei with restriction enzymes. In transgenes containing all three elements of the enhancer region (core, $I\mu$ and MAR), cleavage at a *HindIII* site next to the T7 promoter was three- to ten-fold more efficient than in transgenes containing only the enhancer core and $I\mu$ promoter (data not shown).

Changes in chromatin structure reflect typically an active transcriptional state^{5,16}. Therefore, the accessibility at the T3 and T7 promoters could be affected by endogenous transcription from the $I\mu$ promoter or from other promoters located in genomic sequences flanking the transgenes. We detected transcription across the T7 promoter in the sense or antisense direction in some but not all cell lines examined (Fig. 3). However, we noticed that the active transcriptional state of the transgene did not correlate with factor access at the enhancer-distal T7 promoter, suggesting that transcription is not necessary for accessibility in chromatin. In addition, T7 polymerase access in cell line 12.35, in which a transgene containing the enhancer core is transcribed by endogenous RNA polymerases, suggests that transcription may not be sufficient for chromatin accessibility (Fig. 3).

Enhancers linked to RNA polymerase II promoters, but not by

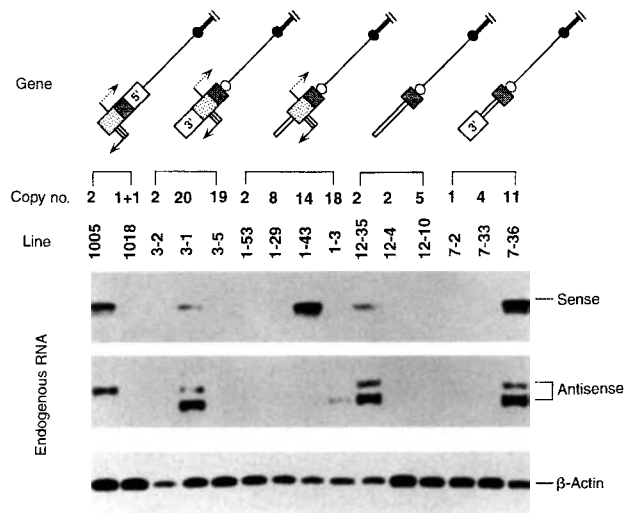


Figure 3 Transcription of transgenes by endogenous RNA polymerases. Transgenic VP1-containing endogenous transcripts that traverse the T7 promoter in a sense or antisense orientation were analysed by S1 nuclease protection assays as described². To control for equal quality and quantity of the RNA preparations, we also examined endogenous mouse β -actin sequences². Transcripts in the sense orientation are likely to originate from the $I\mu$ promoter of transgene copies that are integrated in opposite orientation relative to each other. The slightly slower-migrating DNA fragment protected by antisense transcripts in transgenic line 1005 reflects a sequence heterogeneity caused by a different cloning strategy of that particular gene construct.

themselves, can induce the formation of DNase-I-hypersensitive sites^{2,5}. We examined hypersensitivity to DNase I of the μ enhancer region in both transcriptionally active and inactive transgenic loci (Fig. 4, top panel). As an internal standard for the extent of DNase I digestion, we determined in each sample the DNase I hypersensitivity at the endogenous μ enhancer (Fig. 4, bottom panel). We detected significant DNase I hypersensitivity in chromatin of transgenes containing all three elements of the μ enhancer region and in transgenes containing enhancer core and $I\mu$ promoter, although to hypersensitivity to DNase I of the latter transgenes correlated with the function of the $I\mu$ promoter. By contrast, we found only weak DNase-I hypersensitive sites in the transcriptionally active transgene containing the μ enhancer core (line 12–35).

DNA demethylation has been correlated with changes in chromatin and transcriptional activation of the genes¹⁷. We found that the transgene containing all three elements of the μ enhancer region was ~50% demethylated at a *BstUI* restriction site 0.7 kb away from the enhancer core (Fig. 5). This demethylation was independent of the transcriptional state of the transgene and was observed in pre-B cells but not in non-lymphoid cells (Fig. 5, and data not shown). The extent of DNA demethylation at this restriction site was markedly reduced in the transgene lacking either MAR or the $I\mu$ promoter. By contrast, an *EaeI* restriction site within the μ enhancer core was fully demethylated in all transgenes examined (Fig. 5). Taken together, these data are consistent with the previous observation that demethylation of *in vitro*-methylated Igk gene constructs in transfection assays is dependent on three elements in the intronic enhancer region, but not on the variable region promoter¹⁸.

In conclusion, our data indicate that the μ enhancer core, integrated at different chromosomal positions, can generate only a local accessibility in chromatin but can collaborate with flanking MARs to confer factor access at more distal positions. This extended chromatin accessibility is significantly augmented by sequences within the $I\mu$ promoter region of the enhancer, but is independent of an active transcriptional state. Moreover, a mutation in the Oct

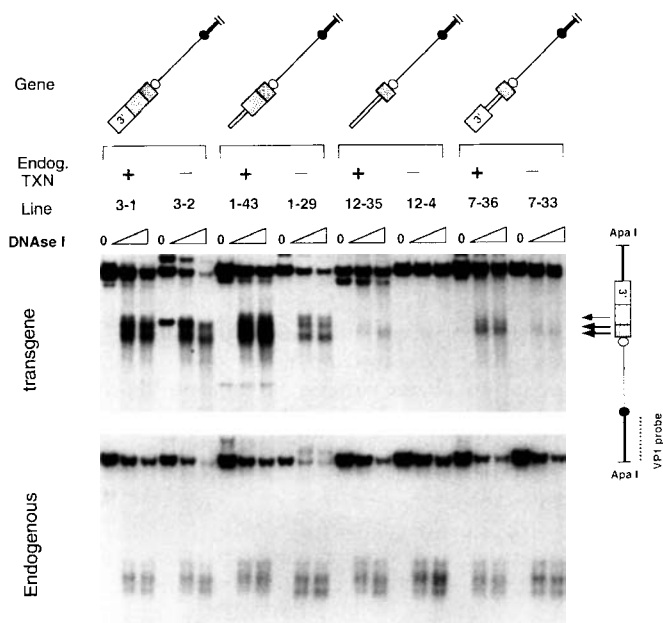


Figure 4 DNase I hypersensitivity of transgenes. Top, mapping of DNase I hypersensitive sites (HS sites) at the transgenic μ enhancer in transcriptionally active (+) and inactive (-) transgenes. Transgenic pre-B cells were incubated with increasing amounts of DNase I. Extracted genomic DNA was adjusted with non-transgenic DNA to compensate for differences in the copy number of the transgene and digested with *Apa*I. Immunoglobulin μ enhancer sequences were visualized by hybridization with a radiolabelled VP1 DNA probe. Arrows on the right-hand side indicate DNase I-generated fragments that map to the enhancer core (bold arrows) and to the $I\mu$ promoter (thin arrow). Bottom, mapping of HS sites in the endogenous μ enhancer. Genomic DNA from the DNase I digestion was cleaved with *Hind*III and analysed by hybridization with a μ DNA probe.

binding site which impairs $I\mu$ promoter function does not affect the enhancer-mediated activation of the VH promoter in transgenic mice⁶. Thus the factor-binding sites in the $I\mu$ promoter region, rather than promoter function, may be important for chromatin accessibility.

Several mechanisms could account for the role of MARs in the enhanced chromatin accessibility and the propagation of factor access to enhancer-distal positions. The finding that MARs, together with the μ enhancer, can induce extended DNA methylation independently of transcription is interesting in the light of the observation that histone H1 associates less efficiently with chromatin containing demethylated DNA¹⁹. Therefore MARs in combination with enhancers and/or promoters may induce localized displacement of histone H1 from chromatin, as inferred *in vitro*²⁰. In addition, acetylation of histones is targeted to transcriptionally active genes and this process may be modulated by MARs^{21–23}. Finally, the immunoglobulin MARs, possibly through binding of proteins such as SatB1 and Bright^{24,25}, may help to recruit chromatin remodelling proteins of the SNF/SWI and NURF families^{26,27}.

The combination of the μ enhancer and flanking MARs may function as a minimal locus control region similar to that of the human CD2 gene, which depends on both enhancer and flanking sequences to overcome heterochromatin-induced position-effect variegation²⁸. Although the chromatin function of locus control regions may be related to the probability of the assembly of a stable multiprotein/enhancer complex in individual cells^{28,29}, our assay cannot address this aspect of the function of the locus control region. A chromatin model of this function is also consistent with the observation that placement of insulator sequences between two divergently transcribed promoters prevents an enhancer from activating the distal, but not the proximal, promoter³⁰. Possibly,

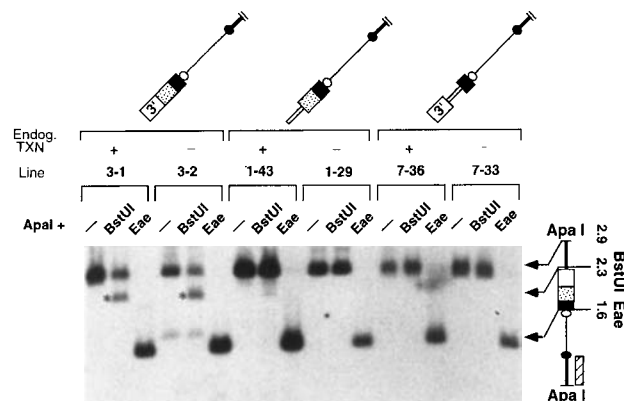


Figure 5 Demethylation of transgenes. Genomic DNA samples from transgenic pre-B cells were digested with *Apa*I alone or in combination with the methylation-sensitive enzymes *Bst*UI or *Eae*I. The map of the restriction enzymes and sizes of DNA fragments, visualized by hybridization with a VP1 probe, are shown on the right. The product of cleavage at the *Bst*UI site is indicated by an asterisk.

insulator sequences impair the propagation of changes in chromatin structure rather than render an enhancer functionally incompetent. Thus, locus control regions may serve two functions in augmenting the activities of polymerase II promoters, the establishment of an extended, accessible chromatin domain, and the stimulation of transcription by interactions with promoter-binding factors.

Methods

Transgene construction. The PEM-T7 and derivative constructs contain a SV40 large-T-antigen (LT) fragment (nucleotides 2,666–3,734) inserted immediately 5' to the T7 promoter in T7VP1 (ref. 2). The boundaries and orientations of the various μ enhancer fragments and 'stuffer' derived from immunoglobulin μ and κ gene^{6,18} sequences were confirmed by sequencing. In the construct lacking the T3 promoter, the polylinker sequence differs from the other μ T3T7 constructs and the T3 DNA probe, generating a longer protected fragment in the S1 nuclease analysis of the antisense transcripts (Fig. 3, line 1005). For microinjection of the transgenic constructs into fertilized C57B16xSJL F₂ mouse eggs, plasmids were released from the vector backbone and the gene inserts isolated. The identification of transgenic fetuses and isolation of transgenic pre-B cell lines have been described².

Preparation of nuclei. Nuclei were prepared by Dounce homogenization of transgenic pre-B cells in nuclei-isolation buffer (0.32 M sucrose, 3 mM CaCl₂, 2 mM magnesium acetate, 0.1 mM EDTA, 10 mM Tris-HCl, pH 7.4, 1 mM dithiothreitol) containing 0.15% N-P40, incubated with bacteriophage RNA polymerases or DNase I, and processed as described². 2×10^7 and 10^8 nuclei were used for analysing RNA polymerase access and DNase I hypersensitivity, respectively.

RNA analysis. The preparation of DNA probes, used for S1 nuclease protection assays of T7-specific transcripts and endogenous sense and antisense transcripts, have been described². For S1 nuclease analysis of T3-specific transcripts, we prepared a DNA probe by extending a 5' [³²P]-end-labelled 21-bp oligonucleotide (5'-GCAGAGCCTGTAGAACCAAC) representing nucleotides 3,673–3,694 of the SV40 genome. For the synthesis of T7-specific transcripts from genomic DNA, we used 10 μ g pre-B cell DNA. For the synthesis of T7- and T3-specific transcripts from chromatin, we incubated 2×10^7 nuclei with 200 units of bacteriophage RNA polymerase.

Received 1 August; accepted 28 November 1996.

- Forrester, W. C., van Genderen, C., Jenuwein, T. & Grosschedl, R. *Science* **265**, 1221–1225 (1994).
- Jenuwein, T., Forrester, W. C., Qiu, R. G. & Grosschedl, R. *Genes Dev.* **7**, 2016–2032 (1993).
- Dillon, N. & Grosfeld, F. *Curr. Opin. Genet. Dev.* **4**, 260–264 (1994).
- Ptashne, M. *Nature* **322**, 697–701 (1986).
- Felsenfeld, G. *Nature* **355**, 219–224 (1992).
- Jenuwein, T. & Grosschedl, R. *Genes Dev.* **5**, 932–943 (1991).
- Adams, J. M. *et al.* *Nature* **318**, 533–538 (1985).
- Lennon, G. G. & Perry, R. P. *Nature* **318**, 475–478 (1985).

9. Cockerill, P. N., Yuen, M. H. & Garrard, W. T. *J. Biol. Chem.* **262**, 5394–5397 (1987).
10. Ernst, P. & Smale, S. *Immunology* **2**, 427–438 (1995).
11. Gasser, S. M. & Laemmli, U. K. *Cell* **46**, 521–530 (1986).
12. Cockerill, P. N. & Garrard, W. T. *Cell* **44**, 273–282 (1986).
13. Aronow, B. J. *et al. Mol. Cell. Biol.* **15**, 1123–1135 (1995).
14. Stief, A., Winter, D. M., Stratling, W. H. & Sippel, A. E. *Nature* **341**, 343–345 (1989).
15. Lorch, Y., LaPointe, J. W. & Kornberg, R. D. *Cell* **49**, 203–210 (1987).
16. Reitman, M., Lee, E., Westphal, H. & Felsenfeld, G. *Mol. Cell. Biol.* **13**, 3990–3998 (1993).
17. Yisraeli, J. & Szyf, M. in *DNA Methylation* (eds Razin, A., Cedar, H. & Riggs, A.) 353–378 (Springer, New York, 1984).
18. Lichtenstein, M., Keini, G., Cedar, H. & Bergman, Y. *Cell* **76**, 913–923 (1994).
19. McArthur, M. & Thomas, J. O. *EMBO J.* **15**, 1705–1714 (1996).
20. Zhao, K., Kas, E., Gonsales, E. & Laemmli, U. K. *EMBO J.* **12**, 3237–3247 (1993).
21. Hebbes, T. R., Clayton, A. L., Thorne, A. W. & Crane-Robinson, F. *author: or C? EMBO J.* **7**, 1395–1402 (1994).
22. Turner, B. M., Birley, A. J. & Lavender, J. *Cell* **69**, 375–384 (1992).
23. Brownell, J. E. & Allis, C. D. *Curr. Opin. Genet. Dev.* **6**, 176–184 (1996).
24. Dickinson, L. A., Joh, T., Kohwi, Y. & Kohwi-Shigematsu, T. *Cell* **70**, 631–645 (1992).
25. Herrscher, R. F. *et al. Genes Dev.* **9**, 3067–3082 (1995).
26. Cullen, K. E., Winston, F. & Carlson, M. *Trends Genet.* **8**, 387–391 (1992).
27. Tsukiyama, T. & Wu, C. *Cell* **83**, 1011–1020 (1995).
28. Ferstenstein, R. *et al. Science* **271**, 1123–1125 (1996).
29. Walters, M. *et al. Genes Dev.* **10**, 185–195 (1996).
30. Cai, H. & Levine, M. *Nature* **376**, 533–536 (1995).

Acknowledgements: T.J. and W.F. contributed equally to this study. We are grateful to G. Felsenfeld for critically reading the manuscript. W.F. is a special fellow of the Leukemia Society of America; L.F. is supported by a postdoctoral fellowship from the Ministerio de Educación y Ciencia of Spain; G.L. is an IMP postdoctoral fellow; T.J. was supported in part by funds from the IMP and the NIH. This work was supported by an NIH grant to R.G.

Correspondence and requests for materials should be addressed to R.G. (e-mail: rgrass@itsa.ucsf.edu).

NMR structure of inactivation gates from mammalian voltage-dependent potassium channels

Christof Antz*, Matthias Geyer†, Bernd Fakler*, Markus K. Schott*, H. Robert Guy‡, Rainer Frank§, Johann Peter Ruppersberg* & Hans Robert Kalbitzer†

* Institute of Physiology, University of Tübingen, Gmelinstrasse 5, 72076 Tübingen, Germany

† Department of Biophysics, Max-Planck-Institute for Medical Research, Jahnstrasse 29, 69120 Heidelberg, Germany

‡ Laboratory of Mathematical Biology, National Institute of Health, 6010 Executive Boulevard, Bethesda, Maryland 20892, USA

§ Zentrum für Molekulare Biologie, Im Neuenheimer Feld, 69120 Heidelberg, Germany

The electrical signalling properties of neurons originate largely from the gating properties of their ion channels. N-type inactivation of voltage-gated potassium (K_v) channels is the best-understood gating transition in ion channels, and occurs by a 'ball-and-chain' type mechanism. In this mechanism an N-terminal domain

(inactivation gate), which is tethered to the cytoplasmic side of the channel protein by a protease-cleavable chain, binds to its receptor at the inner vestibule of the channel, thereby physically blocking the pore^{1,2}. Even when synthesized as a peptide, ball domains restore inactivation in K_v channels whose inactivation domains have been deleted^{2,3}. Using high-resolution nuclear magnetic resonance (NMR) spectroscopy, we analysed the three-dimensional structure of the ball peptides from two rapidly inactivating mammalian K_v channels (Raw3 ($K_v3.4$) and RCK4 ($K_v1.4$)). The inactivation peptide of Raw3 (Raw3-IP) has a compact structure that exposes two phosphorylation sites and allows the formation of an intramolecular disulphide bridge between two spatially close cysteine residues. Raw3-IP exhibits a characteristic surface charge pattern with a positively charged, a hydrophobic, and a negatively charged region. The RCK4 inactivation peptide (RCK4-IP) shows a similar spatial distribution of charged and uncharged regions, but is more flexible and less ordered in its amino-terminal part.

The functional properties of inactivation domains of K_v channels have been analysed in studies combining electrophysiological measurements with site-direct mutagenesis. Both electrostatic and hydrophobic interactions were found to determine the affinity of the ball domain for its binding site in the pore-forming parts of the channel protein^{4,5}. The inactivation domain of Raw3, a mammalian K_v channel expressed in the brain⁶, is inhibited by protein kinase C (PKC) through a phosphorylation of serine residues 15 and 21 (Ser 15 and Ser 21)⁷ and through oxidation of a cysteine residue at position 6 (Cys 6)⁸. Similarly, RCK4, another brain K_v channel⁹, is regulated by oxidation of its cysteine residue 13 (Cys 13)¹⁰.

To identify structure correlates for the functional properties of K_v channel inactivation domains, we analysed the structure of Raw3-IP (30 amino-acid residues in length^{3,8}) in aqueous solution using NMR spectroscopy. The backbone of Raw3-IP exhibits a four-leaf clover-like conformation (Fig. 1b) that leads to a compact structure (Fig. 1c and d), supporting the idea of a 'ball'. All positive charges cluster at one side of the molecule (Fig. 1c), and the two glutamate residues (Glu 27 and Glu 28) form a negatively charged tail close to a hydrophobic region (Fig. 1d). The surface structure at the three functional sites Ser 15, Ser 21 and Cys 6 is shown enlarged in Fig. 1e. The two serine residues targeted by PKC are localized at opposite parts of the molecule. Ser 21 is situated on the hydrophobic hemisphere near the C terminus (Fig. 1e, left), but Ser 15 is localized within the positive cluster (Fig. 1e, centre). Cys 6, whose sulphhydryl group is oxidized in the absence of cytosolic anti-oxidants⁸, is located in close proximity to the side chain of Cys 24 (Fig. 1e, right). All structures shown in Fig. 1 have been determined in the presence of a fivefold excess of the antioxidant dithioerythritol (DTE). To detect possible spontaneous oxidation we performed another set of NMR experiments in the absence of DTE. Two new sets of cysteine cross-peaks were observed in total correlation

Table 1 Structural statistics of Raw3-IP and RCK4-IP

	Raw3-IP	RCK4-IP
Total number of distance constraints	221	486
intra-residual	119	119
sequential ($ i-j = 1$)	66	133
intermediate ($ i-j \leq 4$)	24	58
long range ($ i-j > 4$)	12	0
Dihedral angle constraints	23	33
Distance constraint violations $> 0.5 \text{ \AA}$ (per conformer)	0	0
Dihedral angle constraint violations > 5 (per conformer)	0	0
Average total X-PLOR energy (kJ mol^{-1})	-51.6	-47.2
Average r.m.s.d. from distance constraints	0.22 Å	0.12 Å
Average r.m.s.d. from dihedral constraints	2.7	1.7
Average r.m.s.d. from covalent bonds	0.005 Å	0.004 Å
Average r.m.s.d. from covalent angles	0.71	0.53
R.m.s.d. to the mean for N, C α and C	1.3 Å	0.54 Å*
R.m.s.d. to the mean for all atoms	2.1 Å	1.3 Å*

* Residues 18 to 34.

spectroscopy (TOCSY) spectra, indicating two additional conformers of the peptide. Concomitantly, typical for disulphide bonding, new nuclear Overhauser enhancement (NOE) cross-peaks between the spin systems of the two cysteine residues were observed (NOE connectivities between the amide protons of the two cysteines and the α -proton of one cysteine and the β -protons of the other cysteine). Because the formation of intermolecular disulphide bonds could be excluded by gel electrophoresis (SDS-PAGE without β -mercaptoethanol; data not shown), only the spontaneous formation of an intramolecular disulphide bridge between Cys 6 and Cys 24 could explain the data. As often observed by NMR spectroscopy¹¹, the disulphide bridge in Raw3-IP can exist in two slowly converting isomeric forms. An analysis of the chemical shifts in the different states showed that formation of the disulphide bond causes structural changes only in the vicinity of the two cysteines, with the overall conformation of Raw3-IP remaining unchanged.

For comparison we also analysed the structure of RCK4-IP, an inactivation peptide of 37 amino-acid residues that is derived from

the mammalian K_v channel RCK4 (ref. 3). Similar to Raw3-IP, RCK4-IP binds to non-inactivating RCK4 channels, although its binding affinity is slightly weaker than that of Raw3-IP (ref. 3). The backbone and surface structure of RCK4-IP are shown in Fig. 2, which contrasts the best structure to that of Raw3-IP. RCK4-IP consists of three clearly distinct regions: a highly flexible and disordered N terminus, a well-defined α -helix (Tyr 21 to Ala 34), and a β -turn type II (Pro 18 to Tyr 21) which caps it (Fig. 2b, α -helical region shown in blue, β -turn in green). The two molecules have rather different backbone structure, but both inactivation peptides display charged and hydrophobic surface domains with a similar spatial distribution (Fig. 2c). Moreover, the dipole moments of both ball peptides (calculated with GRASP¹²) are of comparable strength (6.9×10^{-28} Cm for Raw3-IP and 7.8×10^{-28} Cm for RCK4-IP) and direction with respect to these surface domains. This suggests that the two structures have a similar orientation in the transmembrane electrical field.

In conclusion, the inactivation peptides derived from the mam-

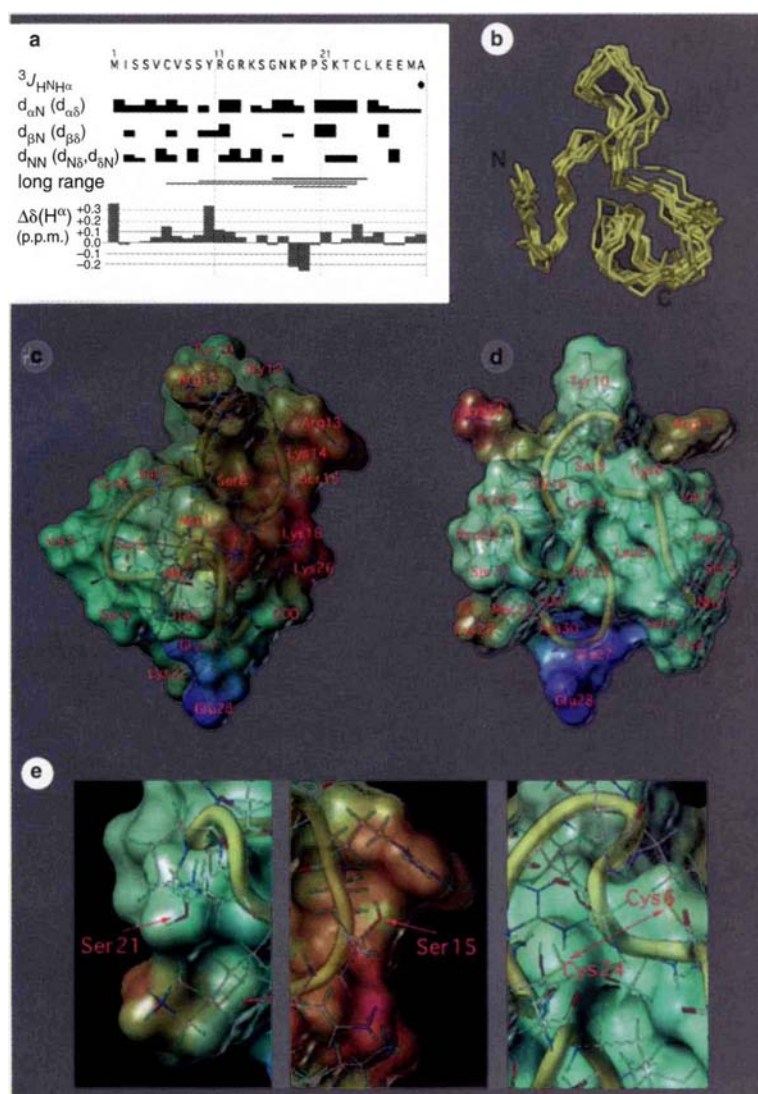


Figure 1 Structure of Raw3-IP. **a**, J -coupling constants ($^3J_{\text{HNH}\alpha}$), NOE connectivities between protons of neighbouring ($d_{\alpha\text{N}}$, $d_{\beta\text{N}}$, d_{NN}) and non-neighbouring amino acids (line thickness reflects volume intensities of NOEs) and secondary chemical shifts of the $\text{H}\alpha$ protons as a function of the amino-acid sequence; the circle refers to a J -coupling constant < 6 Hz. **b**, Superposition of the 8 lowest-energy conformers of Raw3-IP. **c**, **d**, Backbone and surface of the best Raw3-IP structure; the two views are related by a rotation of $\sim 90^\circ$ about the vertical axis.

Backbone atoms are shown as yellow ribbon, N- and C-terminal ends and side chains are as indicated. Colours illustrate the surface charge, which was scaled arbitrarily and ranged from blue (negatively charged) to red (positively charged). **e**, Localization of the residues targeted by phosphorylation (Ser 15 and Ser 21, left, centre) and oxidation (Cys 6, Cys 24). Close proximity of Cys 6 and Cys 24 enables formation of an intramolecular disulphide bridge (arrow, right).

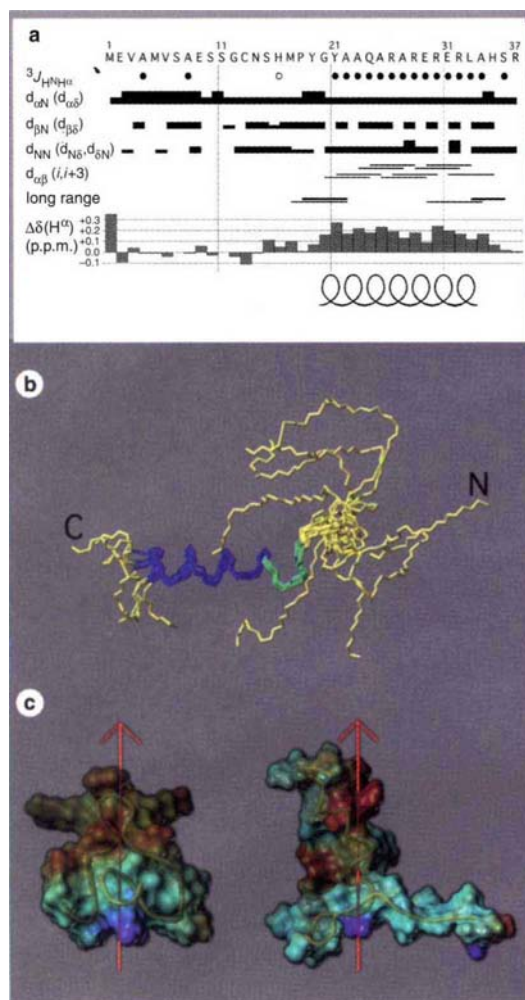


Figure 2 Structure of RCK4-IP. **a**, Plot of J -coupling constants, sequential NOE connectivities, secondary H_α chemical shifts and derived secondary structural elements. Data notation as in Fig. 1a, categories for J -coupling constants are: $J < 6$ Hz (filled circles) and $J > 8$ Hz (open circles). **b**, Display of the 8 lowest-energy structures with an optimal alignment of the backbone heavy atoms N, C_α and C in the α -helical region (amino acids 21–34). **c**, Comparison of the surface structure of Raw3-IP and RCK4-IP, both oriented with respect to their electrical dipole moments. Colours are as in Fig. 1c, d.

malian K_v channels Raw3 and RCK4 display ordered, but clearly different, structures in aqueous solution. This suggests that there is more than one structural class of inactivation peptide. In line with these observations are results obtained from circular dichroism and NMR experiments on the inactivation ball of the *Shaker* B potassium channel. This polypeptide of 20 amino acids was found to be unstructured in aqueous solution^{13,14}. Although the structures shown here for Raw3-IP and RCK4-IP are largely different, both peptides are almost equally potent in blocking open K_v channels. Thus structure-guided experiments are required to establish which parts of these ball domains determine functional interactions with the channel protein. Moreover, the three-dimensional structures of the inactivation balls may be used to validate models of the inactivation process (Fig. 3) and the inner pore of K_v channels. □

Methods

Sample preparation. Raw3-IP and RCK4-IP were synthesized¹⁵ and dissolved in 0.5 ml 99.9% D_2O and in a mixture of H_2O/D_2O (90/10 by value), respectively; final concentrations of the peptides were 2.7 mM, pH 3.5, for

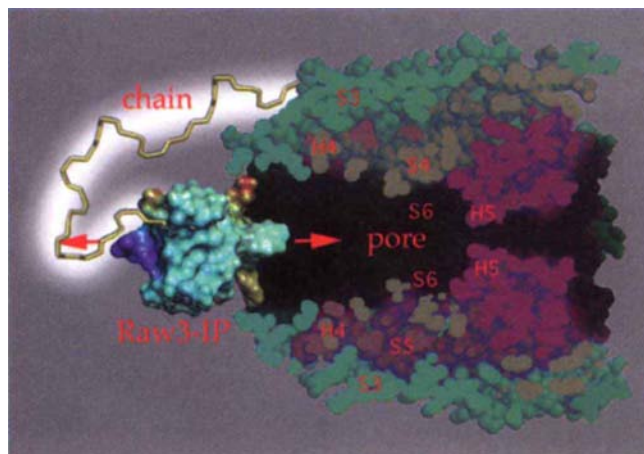


Figure 3 Structural model of the 'ball-and-chain'-type inactivation of K_v channels based on the structure of Raw3-IP and the latest version of a *Shaker* K_v channel²⁴. The model of the channel protein was generated with RasMol 2.5 (Glaxo); colours of the transmembrane segments are as indicated. The chain domain is represented by a yellow ribbon.

Raw3-IP and 2.3 mM, pH 3.5, for RCK4-IP. For verification of structural stability under physiological conditions, both peptides were also dissolved in a physiological buffer solution with and without addition of the reducing agent DTE; pH values were 6.9 with 10 mM DTE and 6.5 without DTE for Raw3-IP, and 6.5 (with and without 4 mM DTE) for RCK4-IP.

NMR spectroscopy. NMR spectra were recorded at 283K with a Bruker AMX-500 spectrometer. 1H chemical shifts were referred to 2,2-dimethyl-2-silapentane-5-sulphonate (DSS). Spin-system identification and sequential assignments were achieved by two-dimensional homonuclear NOESY¹⁶, DQF-COSY¹⁷ and TOCSY¹⁸ experiments. TOCSY experiments were recorded with a mixing time of 60 ms. To identify spin-diffusion effects, NOESY spectra with mixing times between 50 and 400 ms were recorded. $^3J_{HNH\alpha}$ coupling constants were measured in high-resolution DQF-COSY experiments with a digital resolution of 0.66 Hz per point for Raw3-IP, and 0.74 Hz per point for RCK4-IP. **Structure calculations.** Structures were calculated with the program X-PLOR 3.1 (ref. 19). We generated 105 start structures for Raw3-IP and 400 start structures for RCK4-IP from a random field of atomic coordinates with a subsequent regularization and minimization²⁰. The target function that was minimized during simulated annealing²¹ and restrained molecular dynamic simulations comprises quadratic harmonic potential terms for the covalent geometry, angles, torsions and improper terms, and a 6-12 Lennard-Jones or soft quartic potential with a cutoff at 8 Å for the van der Waals repulsion term. An additional conventional Coulomb potential with a cutoff at 8 Å and a dielectrical susceptibility of 3 was used for final refinement. Spectral analysis, peak picking, volume integration and relaxation back-calculation have been performed using the program AURELIA²². Figures were generated with the programs Sybyl (Tripos) and MolMol²³.

Received 11 September; accepted 20 November 1996.

- Hoshi, T., Zagotta, W. N. & Aldrich, R. W. *Science* **250**, 533–538 (1990).
- Zagotta, W. N., Hoshi, T. & Aldrich, R. W. *Science* **250**, 568–571 (1990).
- Ruppersberg, J. P., Frank, R., Pongs, O. & Stocker, M. *Nature* **353**, 657–660 (1991).
- Murrell-Lagnado, R. D. & Aldrich, R. W. *J. Gen. Physiol.* **102**, 949–975 (1993).
- Murrell-Lagnado, R. D. & Aldrich, R. W. *J. Gen. Physiol.* **102**, 977–1003 (1993).
- Schröter, K. H. *et al.* *FEBS Lett.* **278**, 211–216 (1991).
- Covarrubias, M., Wei, A., Salkoff, L. & Vyas, T. B. *Neuron* **13**, 1403–1412 (1994).
- Stephens, G. J. & Robertson, B. *J. Physiol. (Lond.)* **484**, 1–13 (1995).
- Stühmer, W. *et al.* *EMBO J.* **8**, 3235–3244 (1989).
- Ruppersberg, J. P. *et al.* *Nature* **352**, 711–714 (1991).
- Ottig, G., Liepinsh, E. & Wüthrich, K. *Biochemistry* **32**, 3571–3582 (1993).
- Nicholls, A., Sharp, K. A. & Honig, B. *Proteins* **11**, 281–296 (1991).
- Lee, C. W., Aldrich, R. W. & Gierach, L. M. *Biophys. J.* **61**, 379a (1993).
- Fernandez-Ballester, G. *et al.* *Biophys. J.* **68**, 858–865 (1995).
- Frank, R. & Gausepohl, H. *Modern Methods in Protein Chemistry* (de Gruyter, Berlin, 1988).
- Jeener, J. *J. Chem. Phys.* **71**, 4546–4553 (1979).
- Auer, W. P., Bartholdi, E. & Ernst, R. R. *J. Chem. Phys.* **64**, 2229–2246 (1976).
- Davis, D. G. & Bax, A. *J. Am. Chem. Soc.* **107**, 2820–2821 (1985).

19. Brünger, A. T. *X-PLOR Version 3.1* (Yale Univ., New Haven, 1992).
20. Nilges, M., Clore, G. M. & Gronenborn, A. M. *FEBS Lett.* **239**, 129–136 (1988).
21. Kirkpatrick, S., Gelatt, G. C. & Vecchi, M. P. *Science* **220**, 671–680 (1983).
22. Neidig, K. P. *et al.* *J. Biol. NMR* **6**, 255–270 (1995).
23. Koradi, R., Billeter, M. & Wüthrich, K. *J. Mol. Graph.* **14**, 51–55 (1996).
24. Guy, H. R. & Durell, S. R. *Biophys. J.* **62**, 238–250 (1992).

Acknowledgements: We thank K. C. Holmes for support; R. Saffrich for technical assistance; G. J. Busch and L. P. Wollmuth for reading the manuscript.

Correspondence and requests for materials should be addressed to H.R.K. (e-mail: kalbitz@mpimf-heidelberg.mpg.de) or J.P.R. (e-mail: peter.ruppersberg@uni-tuebingen.de). Coordinates have been deposited in the Protein Data Bank (accession nos 1ZTN (for Raw3-IP) and 1ZTO (for RCK4-IP)).

Synergistic effects of substrate-induced conformational changes in phosphoglycerate kinase activation

Bradley E. Bernstein*, Paul A. M. Michels† & Wim G. J. Hol*

* Departments of Biochemistry and Biological Structure, Biomolecular Structure Center and Howard Hughes Medical Institute, University of Washington, Box 357742, Seattle, Washington 98195, USA

† Research Unit for Tropical Diseases, International Institute for Cellular and Molecular Pathology, and Laboratory of Biochemistry, University of Louvain, Brussels, Belgium B-1200

Phosphoglycerate kinase (PGK), a key enzyme in glycolysis, catalyses the transfer of a phosphoryl-group from 1,3-bisphosphoglycerate to ADP to form 3-phosphoglycerate and ATP. Despite extensive kinetic and structural investigations over more than two decades, the conformation assumed by this enzyme during catalysis remained unknown. Here we present the 2.8 Å crystal structure of a ternary complex of PGK from *Trypanosoma brucei*, the causative agent of sleeping sickness. This structure determination relied on a procedure in which fragments containing less than 10% of the scattering mass were successively positioned in the unit cell to obtain phases. The PGK ternary complex exhibits a dramatic closing of the large cleft between the two domains seen in all previous studies^{1–6}, thereby bringing the two ligands, 3-phosphoglycerate and ADP into close proximity. Our results demonstrate that PGK is a hinge-bending enzyme, reveal a novel mechanism in which substrate-induced effects combine synergistically to induce major conformational changes and, to our knowledge, afford the first observation of the PGK active site in a catalytic conformation.

PGK is a monomeric enzyme composed of N- and C-terminal domains connected by a well-conserved hinge region. In all previously reported crystal structures, bound substrates are too far apart to permit catalysis. This has been explained both by a hinge-bending mechanism that can approximate the domains and substrates⁷, and by an alternative binding site for 1,3-bisphosphoglycerate (1,3-BPG)⁸. In the *T. brucei* ternary complex (Fig. 1), the substrates are 6 Å closer together than seen before in any PGK structure and are now well aligned for catalysis. This alignment is accomplished by a 32 degree hinge motion with a positional displacement of up to 27 Å for the ends of the domains, relative to the unliganded horse PGK structure¹. The four subunits in the asymmetric unit of *T. brucei* PGK crystals all show essentially the same domain closure and thereby prove that PGK is a hinge bending enzyme. The dramatic mode of action of this enzyme can now be revealed by taking the structures of horse¹, porcine³, *B. stearothermophilus*⁴, and *T. brucei* PGK as representatives of, respectively, the unliganded, 3-phospho-

TABLE 1 Crystallographic data and refinement statistics

Resolution range (Å)	50.0–2.8
Unique reflections	52,570
Completeness	93.3%
$R_{\text{sym}}[\Sigma I_{\text{obs}} - I_{\text{ave}} / \Sigma I]$	5.5%
Refinement resolution range (Å)	8–2.8
Protein atoms	12,488
Substrate atoms	156
Reflections (test set)	46,091 (3,502)
R factor (R_{free})	23.8% (31.0%)
Average B value (Å ²)	35.0
R.m.s.d. bond lengths (Å)	0.012
R.m.s.d. bond angles (°)	1.68
R.m.s.d. between monomers in the asymmetric unit	
N-terminal domains (Å)	<0.25 (C α)
C-terminal domains (Å)	<0.09 (C α)

glycerate (3-PGA) binary, ADP binary, and ternary PGK complexes. As shown in Fig. 2, binding of 3-PGA alone induces an 8-degree hinge bend with respect to the unliganded form¹. This conformational change is highly localized and consists of a bending of helix 7 (Fig. 1) between residues 194 and 197 (*T. brucei* glycosomal PGK sequence is used throughout) near its N terminus, and a shear motion of helix 14 relative to the N-terminal domain. The additional hinge bending evident in the newly observed closed conformation is attained by a second bend between residues 204 and 207 at the C terminus of helix 7 (Fig. 3a). This latter hinge motion, which accounts for by far the largest part of the 32 degrees of rotation, only occurs once both substrates have bound.

Crucial for understanding PGK is the fact that complete hinge closure is brought about by the synergistic combination of 3-PGA- and Mg-ADP-induced conformational changes (Fig. 3b). 3-PGA binds to the arginine-rich 'basic patch' of the N-terminal domain (Fig. 4) and induces a reorientation of the 3-PGA binding loop, which in turn forces helix 14 to translate in a piston-like motion and invokes the first 8 degree '3-PGA' hinge bend. Helix 14 is now in position to make key interactions with the C-terminal domain once Mg-ADP has also bound. Mg-ADP induces a subtle rearrangement of residues at the N terminus of helix 13 which, when coupled with changes induced by 3-PGA binding, enables a β -sheet to be

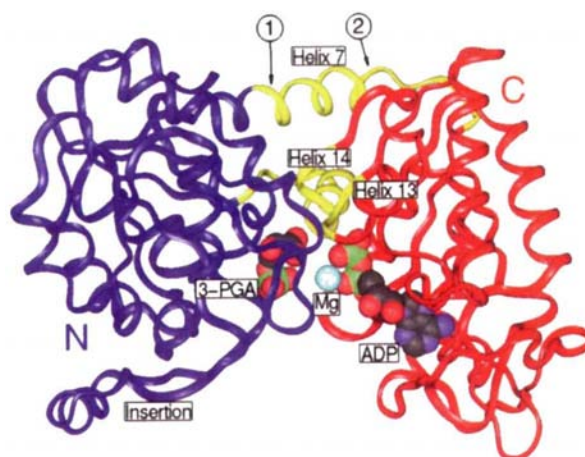
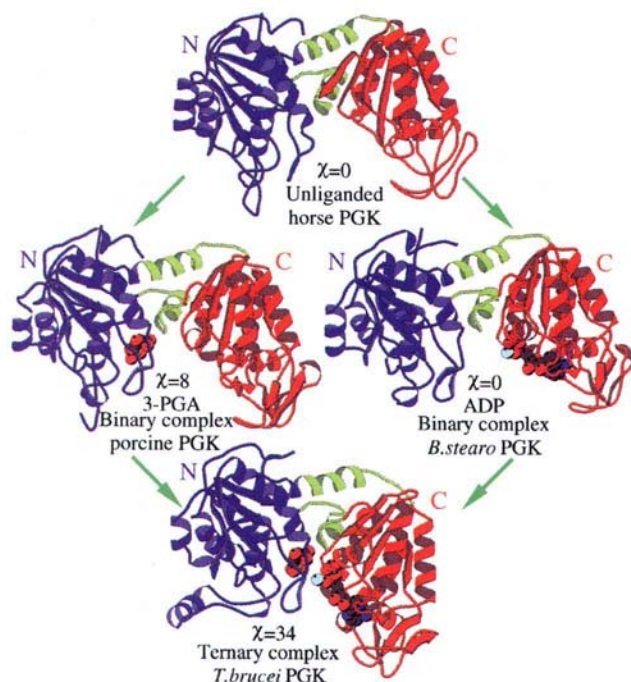


Figure 1 The ternary 3-PGA-Mg-ADP-enzyme complex of *T. brucei* PGK. The N-terminal domain (residues 5–194 and 411–419) is blue, the C-terminal domain (residues 211–394) is red, and the hinge region (residues 195–210 and 395–410) is yellow. Helices '7' (residues 191–205), '13' (residues 376–385) and '14' (residues 397–404) are critical for conformational change and catalysis. Hinge points '1', activated by 3-PGA binding, and '2', activated only once both 3-PGA and Mg-ADP are bound, are indicated by arrows. The insertion contains residues 69–84 and is unique to trypanosomal PGK²⁷.



◀ **Figure 2** Substrate-induced conformational changes in PGK catalysis²³. Four states of ligation are represented by horse¹, *B. stearo*², porcine³ and *T. brucei* PGK, oriented by superimposing their N-terminal domains. Hinge closure (χ) is relative to the unliganded open conformation. Conformational changes in the hinge region (yellow) accommodate closure. The separate N- and C-terminal domains from these PGKs superimpose with r.m.s.d.s for C α atoms of less than 1.7 Å. The enzyme can reach the closed conformation along either pathway since PGK binds its substrates in random order²⁴. Apparently, hinge closure is subject to very subtle differences in substrates as the porcine 3-PGA-MnAMP-PNP⁵ and the yeast 3-PGA-MgAMP-PNP⁶ PGK complexes do not adopt the closed conformation.

completed between residues preceding helix 13 and those preceding helix 14. In this way, hinge closure is achieved and the active conformation stabilized. This precise reorganization, which fixes these helices and the bound substrates in the catalytic conformation, takes place only after both substrates have bound in order to prevent futile hydrolysis of 1,3 BPG or ATP.

The *T. brucei* PGK ternary complex shows the PGK active site in a catalytic conformation. Modelling the substrate 1,3-bisphosphoglycerate shows that the transferring phosphate resides just 4 Å from the Mg-ADP β -phosphate oxygen nucleophile. The

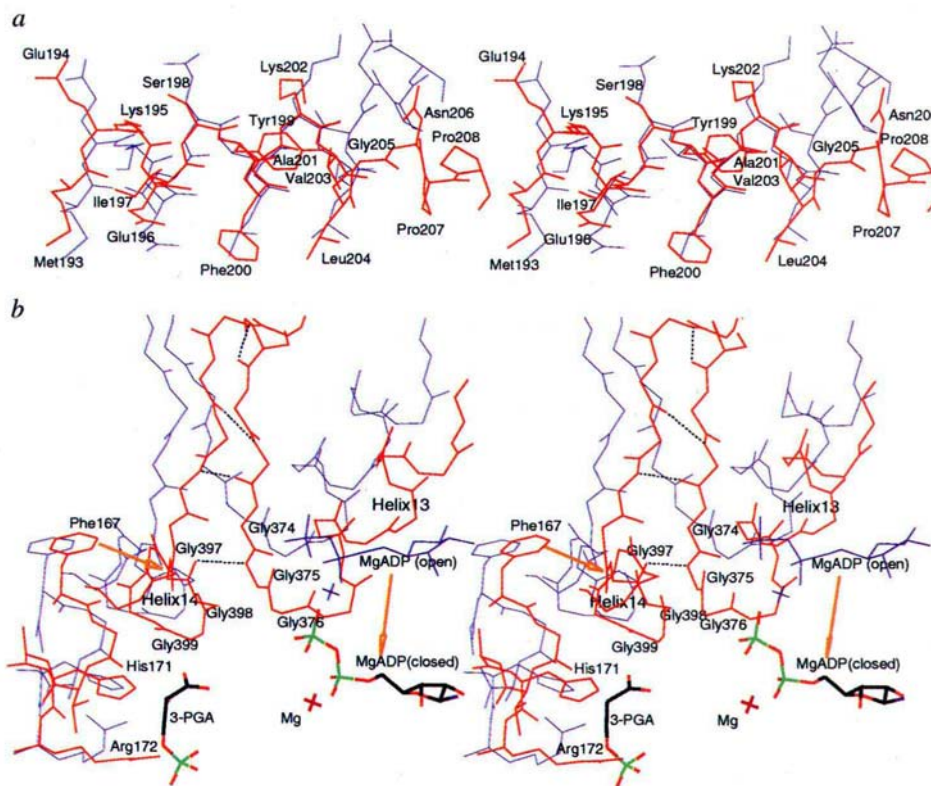


Figure 3 The critical events in PGK hinge closure depicted in stereo view. Open form *B. stearo*² Mg-ADP-PGK¹ is in blue, closed form *T. brucei* ternary complex PGK in red, with helix 7 superimposed in **a**, and N-terminal domains superimposed in **b**. **a**, PGK closure is largely accommodated by a hinge point at the C terminus of helix 7 (residues 204–207). **b**, An interaction between the 3-PGA phosphate group and Arg 172 induces a rearrangement of the 3-PGA binding loop (residues 167–172) which shifts Phe 167 towards helix 14 (left arrow) forcing this helix to translate and inducing the 3-PGA hinge bend. Mg-ADP binds with its β -phosphate anchored to the backbone amides of residues at the N-

terminus of helix 13. As a consequence, residues preceding helix 13 (372–374) are reoriented and form backbone hydrogen bonds with residues preceding helix 14 (395–397). Hinge closure and movement of Mg-ADP towards 3-PGA (right arrow) result. These conformational changes are consistent with mutagenesis^{26,28} and ¹H-NMR studies²⁷. Involved residues are highly conserved in all PGKs²² and include two glycine triplets at the N-termini of helices 13 and 14. These glycine triplets are crucial for substrate binding, conformational change and transition state stabilization (Fig. 5).

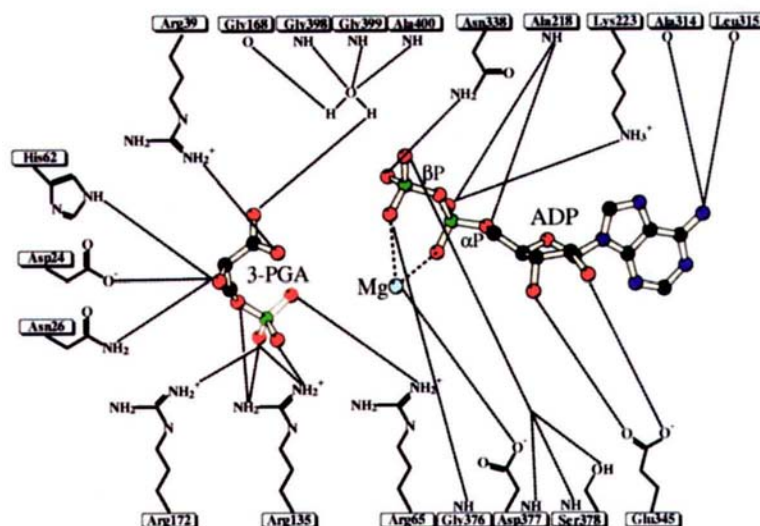


Figure 4 Hydrogen bonds to substrates 3-PGA and Mg-ADP are depicted in schematic. With the exception of an interaction between 3-PGA and the catalytic

Arg 39, which moves upon PGK hinge closure, the binding modes resemble closely those observed in the relevant binary complexes^{3,4}.

relatively short intersubstrate distance and the close proximity of the magnesium ion strongly suggest that phosphoryl transfer in PGK proceeds through an associative mechanism⁸ in which a charged transition state must be stabilized for catalysis⁹. This transition state was modelled as a pentacoordinated phosphoryl group, equidistant from the 3-PGA carboxyl oxygen and a Mg-ADP β -phosphate oxygen (Fig. 5). In the closed form of PGK this phosphoryl group is stabilized very precisely by three key factors. First, one of the three negatively charged phosphoryl oxygens is interacting with the positively charged magnesium. It is likely that in the presence of 1,3-BPG, the divalent magnesium switches from coordinating the α - and β -phosphates of ADP to bridging the β -

phosphate of ADP and the 1-phosphate of 1,3-BPG. Coordination of both substrates by magnesium is consistent with ³¹P NMR analysis of a RhATP-PGA-PGK complex¹⁰, and has been observed in phosphofructokinase¹¹. Second, another phosphoryl oxygen is stabilized by Arg 39, whose guanidinium group also interacts with the 3-PGA carboxylate. This oxygen is also stabilized by the amide of Gly 376 at the N terminus of helix 13, which simultaneously interacts with the ADP β -phosphate. Finally, the third transition-state oxygen interacts with the amide of Gly 399, itself at the N terminus of helix 14. Each of these interactions is important, but we believe that the interaction made by the amide of Gly 399 is the most critical. In the active form of PGK, this amide and the helix-14 dipole are directed precisely at the transitory phosphoryl oxygen (Fig. 5), stabilizing the transition state to a greater extent than either substrate, thereby facilitating enzyme catalysis.

Thus, the *T. brucei* PGK ternary complex illustrates an elegant mechanism in which relatively minor substrate-induced changes combine to induce a major conformational change that serves precisely to arrange substrates and active-site elements for selective catalysis. As will be discussed elsewhere (B.E.B. *et al.*, manuscript in preparation), this structure of trypanosomal PGK will help in the design of selective inhibitors that might be developed into anti-trypanosomal drugs, because the bloodstream form of the sleeping-sickness parasite is entirely dependent on glycolysis for its energy supply¹². The key to success in determining the structure of the *T. brucei* PGK ternary complex was finding the initial correct orientation and position of a search probe containing only 8.4% of the scattering mass of the asymmetric unit. To our knowledge, such a small percentage has never before been used successfully for the determination of phases by molecular replacement¹³. This accomplishment is thus of considerable general relevance for solving the phase problem in protein crystallography as a rapidly enlarging database of structures becomes available for molecular replacement approaches. □

Methods

Expression, purification and crystallization will be described elsewhere (B.E.B. *et al.*, in preparation). Briefly, PGK from the glycosome of *T. brucei* minus a 20-residue glycosome-specific C-terminal signal sequence was expressed in *E. coli* and purified. Dynamic light-scattering analysis suggested that this enzyme formed nonspecific aggregates in solution which could be prevented by the simultaneous addition of 3-PGA and Mg-ADP. In the presence of this substrate

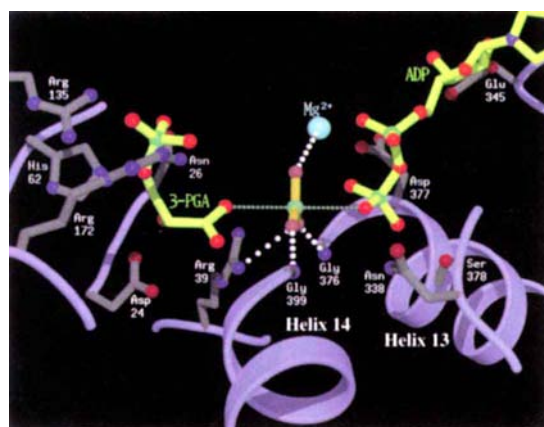


Figure 5 The active site of *T. brucei* PGK with the transition state modelled as a pentacoordinated phosphoryl-group. Substrates are in yellow, key residues in grey. The phosphoryl-group, in orange, resides in a trigonal bipyramidal conformation and is equidistant from the 3-PGA carboxyl oxygen and an ADP β -phosphate oxygen, both at 2.9 Å. This approximate model of the transition state identifies clearly the key elements of PGK catalysis. Interactions with the divalent magnesium ion, Arg 39, and the amides of Gly 376 and Gly 399 predicted to stabilize the transition state are shown. These residues are entirely conserved, and Arg 39 is an absolute requirement for catalysis²⁸ as was also pointed out by May *et al.*⁵. In particular, the NH of Gly 399 is crucial for transition state stabilization. Figure prepared with Raster3D²⁹.

combination, crystals of *T. brucei* PGK were grown by vapour diffusion from solutions containing 2.5 M potassium sodium phosphate at pH 8.0. Data were collected on a single cryopreserved crystal at the Cornell High Energy Synchrotron Source A1 beamline on an ADSC CCD area detector and processed with DENZO and SCALEPACK¹⁴. Crystals belong to space group $P2_12_12_1$ with unit cell dimensions $a = 113.2 \text{ \AA}$, $b = 116.0 \text{ \AA}$, $c = 171.3 \text{ \AA}$; $V_m = 3.05 \text{ \AA}^3$ per dalton¹⁵ for four ternary complexes in the asymmetric unit. Structure determination was non-conventional. Molecular replacement using the known PGK structures²⁻⁴ failed, even though the sequence of *T. brucei* PGK is ~50% identical to these search models. Assuming this was a consequence of significant conformational changes in *T. brucei* PGK, we used the N- and C-terminal domains of *B. stearothermophilus* PGK⁴ as separate search probes. As four independent monomers reside in the asymmetric unit, these fragments each represented less than 10% of the scattering mass, once flexible loops and non-conserved side chains had been removed. Nonetheless, an AmoRe¹⁶ rotation function calculated using data between 8 and 3 Å and a short maximum Patterson vector radius of 20 Å, chosen to minimize inter-domain vectors, identified two top-scoring solutions related by ~90 deg about the z-axis, a relationship that corresponded to a self-rotation function peak. These two orientations also yielded the highest peaks in the translation function. After these N-terminal domains had been fixed, a series of C-terminal domain rotation solutions were tested in a translation search. Two orientations (ranked 16 and 25 in a rotation function using data from 6 to 3 Å) gave the highest peaks. Once oriented and positioned, these four domains packed well in the unit cell and formed two intact subunits with N- and C-terminal domains closely juxtaposed. Based on the orientation of these domains, a new 'closed' search model containing both domains was created that yielded unambiguous molecular replacement solutions for the two additional PGK molecules in the asymmetric unit. After rigid body refinement, an R factor of 50.1% for data from 8 to 3 Å was obtained (Table 1). Phases were improved by iterative solvent flattening and non-crystallographic symmetry averaging¹⁷. The unique 16-residue trypanosome insertion, the hinge region and substrates were gradually built into density with the program O¹⁸ during subsequent cycles of refinement in X-PLOR¹⁹. Non-crystallographic symmetry restraints were imposed separately on the N- and C-terminal domains and on individual components of the hinge region throughout refinement. The Ramachandran plot shows no residues in disallowed regions and the 3D Profile score²⁰ is 224, where 189 or higher is expected. The relative domain orientations of the four *T. brucei* PGK

monomers in the asymmetric unit vary by up to 5 deg. In the crystal structure, two monomers related by a non-crystallographic two-fold axis pack together to form a dimer which buries 1,500 Å² of surface area. The asymmetric unit contains two such dimers, related by an 85° rotation about the z-axis. The dimer is not likely to be physiologically relevant as *T. brucei* PGK²¹ and the *T. brucei* PGK ternary complex (B.E.B. and W.G.J.H., unpublished dynamic light scattering data) are monomeric in solution.

Received 2 October; accepted 18 November 1996.

1. Blake, C. C. F. & Evans, P. R. *J. Mol. Biol.* **84**, 585–601 (1974).
2. Watson, H. C. *et al.* *EMBO J.* **1**, 1635–1640 (1982).
3. Harlos, K., Vas, M. & Blake, C. C. F. *Proteins* **12**, 133–144 (1992).
4. Davies, G. J. *et al.* *Acta Crystallogr. D* **50**, 202–209 (1994).
5. May, A., Vas, M., Harlos, K. & Blake, C. C. F. *Proteins* **24**, 292–303 (1996).
6. McPhillips, T. M., Hsu, B. T., Sherman, M. A., Mas, M. T. & Rees, D. C. *Biochemistry* **35**, 4118–4127 (1996).
7. Banks, R. D. *et al.* *Nature* **279**, 773–777 (1979).
8. Mildvan, A. S. *Phil. Trans. R. Soc. Lond. B* **293**, 65–74 (1981).
9. Knowles, J. R. *Annu. Rev. Biochem.* **49**, 877–919 (1980).
10. Pappu, K. M., Gregory, J. D. & Serpersu, E. H. *Arch. Biochem. Biophys.* **311**, 503–508 (1994).
11. Shirakihara, Y. & Evans, P. R. *J. Mol. Biol.* **204**, 973–994 (1988).
12. Oppendoerf, F. R. *Annu. Rev. Microbiol.* **41**, 127–151 (1987).
13. Oh, B. *Acta Crystallogr. D* **51**, 140–144 (1995).
14. Otwinowski, Z. in *Data Collection and Processing* (eds Sawyer, L., Isaacs, N. & Burley, S.) 56–62 (SERC Daresbury Laboratory, UK, 1993).
15. Matthews, B. W. *J. Mol. Biol.* **33**, 491–497 (1968).
16. Navaza, J. *Acta Crystallogr. A* **50**, 157–163 (1994).
17. CCP4 *Acta Crystallogr. D* **50**, 760–763 (1994).
18. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. *Acta Crystallogr. A* **47**, 110–119 (1991).
19. Brünger, A. T. *X-PLOR Version 3.1* (Yale Univ. Press, New Haven, CT, 1992).
20. Luthy, R., Bowie, J. U. & Eisenberg, D. *Nature* **356**, 83–85 (1992).
21. Misst, O., Bos, O. J. M. & Oppendoerf, F. R. *Eur. J. Biochem.* **157**, 441–453 (1986).
22. Fothergill-Gilmore, L. A. & Michels, P. A. M. *Progr. Biophys. Molec. Biol.* **59**, 105–235 (1993).
23. Kraulis, P. J. *J. Appl. Crystallogr.* **24**, 946–950 (1991).
24. Scopes, R. K. *Eur. J. Biochem.* **85**, 503–516 (1978).
25. Tompa, P., Hong, P. T. & Vas, M. *Eur. J. Biochem.* **154**, 643–649 (1986).
26. Walker, P. A., Littlechild, J. A., Hall, L. & Watson, H. C. *Eur. J. Biochem.* **183**, 49–55 (1989).
27. Graham, H. C. & Williams, R. J. P. *Eur. J. Biochem.* **197**, 81–91 (1991).
28. Sherman, M. A., Fairbrother, W. J. & Mas, M. T. *Prot. Sci.* **1**, 752–760 (1992).
29. Merritt, E. A. & Murphy, M. E. P. *Acta Crystallogr. D* **50**, 869–873 (1994).

Acknowledgements: Data were collected at the Cornell High Energy Synchrotron Source. Coordinates of porcine PGK were kindly provided by C. C. F. Blake. Equipment funds received from the Murdock Charitable Trust and the School of Medicine of the University of Washington are gratefully acknowledged. B.E.B. was supported by the Poncin Scholarship Fund. We thank S. Allert, H. Kim, P. Lynch, F. Merritt, M. Redinbo, P. Petra, G. Rudenko, S. Sarfaty, B. Stoddard, S. Turley, F. van den Akker, J. Verburg, C. Verlinde and V. Yee for assistance.

Correspondence and requests for materials should be addressed to W.G.J.H. (e-mail: hol@gouda.bmsc.washington.edu).

KNOW YOUR COPYRIGHT RIGHTS RESPECT OURS

The publication you are reading is protected by copyright law.

Photocopying copyright material without permission is no different from stealing a magazine from a newsagent, only it doesn't seem like theft.

If you take photocopies from books, magazines and periodicals at work your employer should be licensed with CLA.

Make sure you are protected by a photocopying licence.



The Copyright Licensing Agency Limited
90 Tottenham Court Road, London W1P 0LP
Telephone: 0171 436 5931 Fax: 0171 436 3986